View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Wits Institutional Repository on DSPACE

Finding Statistical Models using Psychometric Tests, Matric Results, and Biographical Data to Predict Academic Success at a South African University

Johanna Wilhelmina Breytenbach

A dissertation submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg in fulfilment of the requirements for the degree of Master of Science

Johannesburg, 2008

Declaration

I declare that this dissertation is my own, unaided work. It is being submitted for the degree of Master of Science in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in and other University.

28th day of July 2008

Abstract

Tertiary education is expensive for the individual, the family, the institution and the country. The aim of this study was to find models that could predict academic success, which is seen in this study as the completion of a specific degree in the minimum prescribed time.

The reliability and construct validity of the subtests of several psychometric tests used were examined. The subtests that were found reliable and construct valid, as well as biographical data and academic history data, were used as independent variables in model fitting procedures to find models to predict academic success.

The SAT 78, GSAT, SSHA, PHSF and 19 FII remain reliable instruments. The SAT 78, PHSF, and 19FII were not found to be construct valid on the sample examined. The GSAT and SSHA were found to be possibly construct valid.

Four different models for each of BCom, BPharm, BA and BSc students were found to predict academic success. The CHAID procedure, which was initially used as an exploratory method selected matric results as the most significant predictor for the BCom, BPharm and BSc degree students. Matric results in combination with other predictors were selected by both the stepwise logistic regression and stepwise predictive discriminant procedures as the best predictors of academic success for all four models at the North-West University's Potchefstroom Campus.

Acknowledgements
I want to thank everyone who helped me with this study.

Peter Fridjhon, my supervisor: Peter, thank you so much for sharing your expertise with me, for motivating me and for your dedication to my study. Thank you for making this experience one of the highlights of my life.

Faans Steyn, my co-supervisor: Faans, thank you for al your effort and time. Working with you over the years learned me so much - not only about statistics but about life.

Jacky Galpin for giving me this opportunity.

Suria Ellis, **Gerhard Koekemoer**, **Jan du Plessis,** and **Christa Labuschage**, my colleagues for all your help.

Leonard Santana, **James Allison, Tiny du Toit,** and **Stefan Jansen van Vuuren** for your support.

Ria Terblanche and her colleagues at the Ferdinand Postma Library.

The **North West University** for using their data.

Nic Kotze for giving permission to use the psychometric tests' data.

Willem Pienaar for allowing to me to use the biographical data.

Annetjie de Waal for supplying matric results.

Amanda Lourens for sharing articles and information.

Johann Schepers for sharing your thoughts and knowledge with me.

Gordon Kass for assistance with CHAID.

Leone Wolmarans for your neat mathtype equations.

Jaco Breytenbach (Snr) for the language and technical editing of this dissertation.

Mike Breytenbach for computer assistance.

Jaco Breytenbach (Jnr) for language assistance.

Mynhardt and **Anke Breytenbach** for your support in many ways.

Sinah Plaatjie and **Emily Molebatse** for all your help.

Dedication

This dissertation is dedicated to:

My husband **Jaco**,

my children **Mike**, **Mynhardt**, **Anke,** and **Jaco**,

my mother **Annatjie**,

and my sister **Marie**.

CONTENTS

List of tables

List of figures

Chapter 1

1. Introduction

1.1. Background

Education of people all over the world is politically and economically a controversial field. Many people do not have the opportunity to receive tertiary education. One of the primary reasons for this is that a quality education is very expensive. Therefore, it is essential for the student, the university, and the state to spend money on education efficiently.

In 2002 a consortium of U.S. educators visited South Africa with the goal of ascertaining which issues were the most pressing for government and higher education. The one theme that consistently arose in interviews around South Africa was the lack of research and data to support decision-making in areas of academic success or retention and attrition of students. One of the main purposes of an institution of higher education is to produce graduates who can contribute to the socio-economic well-being of society. In South Africa there is also great concern about large numbers of students who do not complete their degrees (Lourens, 2006).

Research on academic success or retention studies is more than 100 years old. Retention refers to a student's failure to complete a particular course in a particular time. Information on factors influencing academic success or retention rates in South Africa is limited. The National Plan for Higher Education (Department of Education, 2001) indicated that the reasons for the decline in pass rates are not clear and require investigation. It also states that the focus of higher education institutions should be to increase the number of graduates through improving the efficiency of the higher education system. Until September 2006 there were no national figures available to assess the extent of pass rates in universities in South Africa. The Minister of Education, Ms Naledi Pandor, recently released figures about drop-out rates. Half the country's undergraduate students drop out or are excluded without completing their degrees or diplomas. These figures are not just for historically black universities but include historically white universities as well. The figures are drawn from a department of education "cohort study" of students who first entered undergraduate programmes in 2000. They are seen as the most reliable to date. The student cohort was tracked at each tertiary institution for five years, up to the end of 2004 (Macfarlane, 2006). Retention studies are a complex issue and not merely a compilation of figures. They also

entail an analysis of the causes of low pass rates and the implementation of measures to counter the drop-out of students (Lourens, 2006).

Tinto (1975) had developed an integration-commitment model of attrition. It was later modified by Pascarella and Terenzini (1983) and has been used repeatedly in past research. According to this model perseverance is strongly related to a student's (1) level of academic and social integration with an institution, (2) commitment to earning a degree (goal commitment), and (3) commitment to an institution (institutional commitment).

Liu (2000) stated that commonality between integration and satisfaction is crucial to the success of academic performance and persistence and that students' satisfaction is highly related to student success.

It is important for administrators to understand the unique combination of factors contributing to student attrition at their institutions (Lourens, 2006).

Every year the North-West University spends a considerable amount of time and money on psychometric tests to guide students registering for appropriate courses at their Potchefstroom Campus (Engelbrecht, 1999; Kotze, 1994). In the case of Pharmacy students the tests are used in selection procedures. In addition, certain biographical data as well as every student's matric results are captured. Thus far these data sets have not been used to establish whether predictions of academic success over a 3 to 4 year period (which is the general duration of the various bachelor's degrees) could be made. If statistical models using these variables could be found to predict what type of qualities a student should have to be successful in academic achievement, it would be of great value.

1.2. Aim of the Study

The aim of this study was to undertake a statistical investigation to find the best statistical models to predict academic success in terms of graduation. The study aimed to find these models for the Potchefstroom Campus of the North-West University. Available reliable and valid psychometric subtests, biographical data as well as matric results of students were used. For the purpose of this study the response (dependent) variable is academic success. From a statistical viewpoint, a student is considered as an academic success if his or her degree was completed in the minimum prescribed time, and as a failure if not.

This aim was twofold, namely firstly, to ensure the reliability and validity of the subtests of the psychometric tests, and secondly, to use these reliable and valid subtests together with some available biographical and academic data for model fitting purposes.

1.2.1. Psychometric Tests' Reliability and Construct Validity

The psychometric tests used in this study were the following: the Senior Aptitude Test 1978 edition (SAT 78), the General Scholastic Aptitude Test (GSAT), the Brown-Holtzman Survey of Study Habits and Attitude (SSHA), the Personal Home, Social, and Formal Relations questionnaire (PHSF), and the 19 Field Interest Inventory (19 FII).

The available data of these psychometric tests of the North-West University had been used by four researchers to date, namely Kotze (1994), Engelbrecht (1999), Volschenk (1997), and van Wyk (1988). In these studies the reliability and construct validity of these tests were not determined on the study samples which they have used. Given the diversity of participants across studies, researchers using psychometric tests should provide reliability coefficients on the scores for the data analysed (Pedhazur & Schmelkin, 1991). The same argument holds for the construct validity of the tests (Nunnally & Bernstein, 1994). Little evidence could be found in the literature of any confirmation of these tests' reliability or construct validity on recent data in South Africa (Foxcroft, Paterson, le Roux & Herbst, 2004).

The world and especially South Africa has changed materially in the years since these tests were constructed and standardised. If children had to be seen and not heard in the past, educators do not agree about this anymore. Use of computers and electronic gadgets is common amongst learners and influences their views on life and learning. In the light of this it is meaningful to ask the question of how reliable and construct valid the SAT 78, GSAT, PHSF, SSHA, and 19 FII are on a group of students many years after their initial standardisation.

1.2.2. Finding Models to Predict Academic Success

After the validation of the reliability and construct validity of the subtests of the psychometric tests the reliable and construct valid subtests as well as gender and matric results could then be used to try to find models to predict academic success as described in Section 1.1.

1.3. Research Questions

This study addressed the following research questions:

1. Are the SAT 78, GSAT, SSHA, PHSF, and 19 FII reliable instruments for the study sample and, if so, how does the reliability of these instruments compare with their reliability at the time of their standardisation?

2. Are the SAT 78, GSAT, SSHA, PHSF, and 19 FII construct valid instruments for the study sample and, if so, how does the construct validity of these instruments compare with the construct validity at the time of their standardisation?

3. Which of the available reliable and construct valid predictors are the best at predicting academic success for BCom, BPharm, BA, and BSc students, respectively?

4. Are there valid models which can adequately predict academic success for each of the BCom, BPharm, BA, and BSc students?

1.4. Research Methods

This investigation consists of a theoretical component and an empirical component. The theoretical part covers the possible techniques and methods to assist the empirical component. Information on the psychometric tests is given in Chapter 2 and that of the statistical techniques in Chapter 3.

The empirical component of this study had been done on the available data to establish reliability and validity of the psychometric tests. Data of the classes of 2003, 2004, 2005, 2006, and 2007 were used. For model fitting purposes the data of the classes of 2003 and 2004 were used for the BCom, BPharm, BA, and BSc groups.

Applicants for the BPharm degree were required by the university to do the GSAT, SSHA, PHSF, and 19 FII for selection purposes. For any other students the psychometric tests were not compulsory, but if they chose to do them they were required to complete the SAT 78, SSHA, PHSF, and 19 FII. All the testees were students of, or applicants to this university where students are predominantly white Afrikaans speaking, and from formerly advantaged communities. From an academic point of view the respondents were selected groups, since all of them had either passed grade nine (in the case of the applicants for pharmacy) or were already admitted to enter the university. By taking into consideration that no random

selection of students had been made at any time when gathering the data, effect sizes were used to describe relationships and no inferential statistics were used.

The statistical techniques and methods used are discussed in Chapter 4. In Chapter 5 the results of the psychometric tests' reliability and validity are given and in Chapter 6 results of the model fitting procedures are given. In Chapter 7 conclusions, limitations and recommendations are discussed.

Chapter 2

2. Literature Study: Psychometric Testing

In this chapter literature regarding the reliability and validity of psychometric tests is discussed. The psychometric tests used in this study are also reviewed.

2.1. Introduction

A psychological test is an instrument that indicates how much the participant has of the quality the test measures (Anastasi & Urbina, 1997). Psychological tests are like tests in other sciences, for example, a pathologist who measures a patient's white blood cell count or a dietician who measures a person's mass before and after a diet. The psychologist proceeds in much the same way when making observations about an individual's behaviour or aptitudes like arithmetic or reading skills.

2.2. Historical Background

The earliest evidence of standardised testing based on merit comes from 1000 BC when the Chinese introduced written tests to help fill civil service positions. The first main objective of psychological tests was that of measuring intelligence (Aiken & Groth-Marnat, 2006).

In the early 19th century, there was strong interest in classifying types of mental disabilities. Francis Galton proposed the development of measures of central tendency and variability to summarize data and also developed the concept of correlation. In 1890 James McKeen Cattell, a student of Galton, was the first person to use the term "mental test". He developed a set of tests that were able to predict a child's scholastic achievement. Cattell's goals were related to his desire to strengthen psychology's scientific credentials. Karl Pearson, also a student of Galton, developed several statistical measures and techniques still being used today in modern statistics, such as the standard deviation and the normal curve. His most well known statistical concept is the product moment correlation coefficient, or Pearson's r (Aiken & Groth-Marnat, 2006; Anastasi & Urbina, 1997; Nunnally & Bernstein, 1994).

Alfred Binet was the first person to formulate a test for children with mental challenges. He spoke strongly about the nature-nurture controversy, believing that intelligence could be nurtured, and was not simply the product of nature. Binet developed cognitive exercises called "mental orthopaedics" to increase the intelligence level of children. David Wechsler, a

student of Pearson, developed two widely used intelligence scales: the Wechsler Adult Intelligence Scale (WAIS) and the Wechsler Intelligence Scale for Children (WISC). Wechsler defined intelligence as “the global capacity to act purposefully, to think rationally and to deal effectively with the environment” (Aiken & Groth-Marnat, 2006).

Anne Anastasi, one of the best-known psychologists in the field of testing, states that psychological tests are tools that can be instruments of good or harm, depending on how they are used. She defines a test as an "objective" and "standardised" measure of a sample of behaviour (Anastasi & Urbina, 1997).

2.3. Reliability

2.3.1. Introduction

The reliability of a test refers to the consistency of scores obtained by the same persons when they are re-examined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions (Anastasi & Urbina, 1997).

Various theories of reliability have been formulated over the years. The focus will be on classical test theory because it served as a departure point for most of the other theories concerning the estimation of reliability.

In 1904 Spearman proposed the true-score model, what has become known as classical test theory. According to this model, any observed score consists of two components, namely a true component and an error component, say

$$X = T + E \tag{2.1}$$

where X is the imperfect, observed score, T is the true score, and E is the random error (Pedhazur & Schmelkin, 1991). Although theoretically meaningful, (2.1) contains two unknowns and cannot be solved without further assumptions. Thus in classical test theory it is assumed that the traits measured are constant and the measurement errors random.

Being random, the mean of measurement errors over many repeated measurements is expected to be zero. That is

$$E(E) = 0 \tag{2.2}$$

and the true score is thus equal to the expected observed scores over a number of repeated measurements (2), namely

$$T = E(X) \tag{2.3}$$

Let

$$x = t + \varepsilon \tag{2.4}$$

where $x = X - \bar{X}$, $t = T - \bar{T}$, and $\varepsilon = E - \bar{E}$ (the raw scores minus their respective means).

Since an observed score is a composite of true and error scores the variance of x is

$$\begin{aligned} \sigma_x^2 &= \sigma_{(t+\varepsilon)}^2 \\ &= \sigma_t^2 + 2\sigma_{t\varepsilon} + \sigma_\varepsilon^2, \end{aligned} \tag{2.5}$$

where σ_t^2 = variance of true scores, σ_ε^2 = variance of errors and $\sigma_{t\varepsilon}$ covariance of true and error scores. However, because the true and error scores are independent, $\sigma_{t\varepsilon} = 0$ and therefore

$$\sigma_x^2 = \sigma_t^2 + \sigma_\varepsilon^2. \tag{2.6}$$

Further is the correlation between observed scores (i.e. $t + \varepsilon$) and true scores (t)

$$\begin{aligned} r_{xt} &= \frac{\sum (t+\varepsilon) t}{N\sigma_x\sigma_t} \\ &= \frac{\sum t^2 + \sum t\varepsilon}{N\sigma_x\sigma_t} \\ &= \frac{\sigma_t^2 + \sigma_{t\varepsilon}}{\sigma_t\sigma_x} \\ &= \frac{\sigma_t^2}{\sigma_t\sigma_x} = \frac{\sigma_t}{\sigma_x}. \end{aligned} \tag{2.7}$$

This means that the correlation between observed scores and true scores is equal to the ratio of the standard deviation of true scores to the standard deviation of observed scores (Anastasi & Urbina, 1997; Nunnally & Bernstein, 1994).

The square of the correlation coefficient (r) indicates the proportion of variance shared by variables being correlated. In this context the squared correlation indicates the proportion of variance in the scores that is due to true differences among the people being measured.

This proportion of variance is also denoted by r_{xx}:

$$r_{xx} = r_{xt}^2 = \frac{\sigma_t^2}{\sigma_x^2}, \tag{2.8}$$

where r_{xx} is the reliability of measure X. According to Pedhazur and Schmelkin (1991) the definition then of the reliability of a measure in mathematical terms is the ratio of true score variance to observed score variance.

By substituting (2.6) in r_{xx} then

$$\begin{aligned} r_{xx} &= \frac{\sigma_x^2 - \sigma_\varepsilon^2}{\sigma_x^2} \\ &= 1 - \frac{\sigma_\varepsilon^2}{\sigma_x^2}. \end{aligned} \tag{2.9}$$

The values of r_{xx} can range from 0 to 1. It is 1 when all the observed variance is due to true score variance. That is when there are no random errors of variance. At the other extreme when all the observed variance is due to random errors the reliability is 0. The reliability coefficient is interpreted as the proportion of systematic (i.e. true score) variance in the observed scores. For example, r_{xx} = 0.75 means that 75% of the variance of the observed scores is systematic and 25% is the proportion of variance due to random errors. As a result of the fact that according to (2.8) the reliability coefficient is actually a squared correlation coefficient which is always sample specific, one must avoid speaking of the reliability of a given instrument, without specifying the population from which the sample was taken.

Unfortunately the equations developed above cannot be used for determining reliability because they include an element that refers to an unobservable true score variance. The reliability therefore needs to be estimated from the sample's observed scores X.

2.3.2. Estimation

Reliability can be regarded as a theory of errors. As a result of the fact that measurement errors may come from different sources, as seen in Section 2.3.1, reliability estimates will differ. Therefore, it is important, when reporting reliability estimates to include information about procedures used, to inform the reader about the sources of error. The three most used approaches to the estimation of reliability are discussed below.

2.3.2.1. Test-retest

According to this approach, a group of people is measured twice, using the same measure, and the two sets of scores thus obtained, are correlated. The underlying assumption is that the correlation between the two sets of variables is due to the underlying unobservable true scores that are constant and the correlation will not be perfect as a result of the random errors of measurement. Many problems can occur, for example, if participants remember the test items, a carry-over effect tends to inflate the estimate of the measure's reliability (Anastasi & Urbina, 1997; Nunnally & Bernstein, 1994; Pedhazur & Schmelkin, 1991). To counteract this effect the time between the two measures could be increased, with the implication that a low test-retest correlation is the result of true changes in the measured characteristic of the individuals.

2.3.2.2. Equivalent Forms

With this method two different forms of a measure are designed to measure the same phenomenon. The underlying assumption here is that the two forms should be parallel, that is that the two forms have identical true scores and equal error variance. In mathematical terms, two measures X_1 and X_2, are said to be parallel if

$$X_1 = T + E_1$$
$$X_2 = T + E_2$$
$$\sigma^2_{\varepsilon_1} = \sigma^2_{\varepsilon_2}. \qquad (2.10)$$

The correlation between the two forms is then taken as an estimate of the reliability of either of them. One of the problems is that it is nearly impossible to construct such forms and to

determine then if they are really parallel (Anastasi & Urbina, 1997; Nunnally & Bernstein, 1994; Pedhazur & Schmelkin, 1991).

2.3.2.3. Internal Consistency

Because of the previous noted problems, a conception of reliability based on a single administration of a measure has been developed, called internal consistency. In this case, the measures must be composed of multiple items based on items measuring the same phenomenon. The underlying idea (assumption) is then that responses to items in a composite measure are expected to be internally consistent (Anastasi & Urbina, 1997; Nunnally & Bernstein, 1994; Pedhazur & Schmelkin, 1991).

Split-half Reliability Estimates

The first measure of the internal consistency approach to estimation of reliability is split-half reliability estimates. This method is a variation on the alternate forms estimate, each half is treated as if it were an alternate form of the other. This means then that this correlation is based on a measure half as long as the original measure. Spearman and Brown developed a formula, named the Spearman-Brown formula, based intuitively on the expectation that by increasing the size of an instrument, its reliability would increase and by decreasing its size its reliability would decrease. The validity of this formula is based on the assumption that parts added or subtracted from the measure are strictly parallel to the original measure. In mathematical terms the formula is written as

$$r_{kk} = \frac{kr_{xx}}{1+(k-1)r_{xx}}, \tag{2.11}$$

where k is the factor by which the instrument is increased or decreased, r_{xx} is the reliability of the existing measure and r_{kk} is the estimated reliability of an instrument k times longer or shorter than the original one.

A measure may be split in half in many different ways. However, the assumption is that the two halves must be parallel, which is always difficult to assure (Anastasi & Urbina, 1997; Nunnally & Bernstein, 1994; Pedhazur & Schmelkin, 1991). For example, suppose that a measure of achievement is used in which the items are ordered in ascending difficulties. By splitting the measure into two halves by placing the first half in a group and the second in another group (namely the first-second approach) these two halves will yield poor reliability as a result of the fact that the halves are obviously not parallel.

Coefficient α

The internal-consistency approach to the estimation of reliability is based on the fact or assumption that the items or sub-scales of the instrument measure the same thing, i.e. the items or sub-scales must be homogeneous. Various theoretical formulations regarding approaches to internal consistency estimation have been advanced. Although they have different assumptions and uses of analytical approaches, they all arrive at essentially the same estimates of reliability. The most widely used measure of internal consistency is the alpha coefficient, also referred to as Cronbach's alpha:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum \sigma_i^2}{\sigma_x^2}\right], \tag{2.12}$$

where k is the number of items, $\sum \sigma_i^2$ is the sum of the variances of the items and σ_x^2 the variance of the total score, i.e., the composite score (Pedhazur & Schmelkin, 1991).

Furthermore the variance of a composite score equals the sum of the variances of its components plus twice the sum of the covariances of all possible pairs of its components:

$$\sigma_x^2 = \sum \sigma_i^2 + 2\sum \sigma_{ij}, \tag{2.13}$$

where σ_{ij} is the covariance of items i and j $(i \neq j)$.

Furthermore, if there are two variables X and Y their correlation is defined as

$$r_{xy} = \frac{\sum xy}{N\sigma_x\sigma_y} \tag{2.14}$$

where $\sum xy$ is the sum of the products of the deviations of X and Y from their means, N is the number of observations, and σ_x and σ_y are the standard deviations of X and Y, respectively. Their covariance can be expressed as

$$\sigma_{xy} = \frac{\sum xy}{N}, \tag{2.15}$$

where σ_{xy} is the covariance between X and Y, and the other terms are as defined under (2.6).

Using (2.14) the covariance can be expressed as

$$\sigma_{xy} = r_{xy}\sigma_x\sigma_y. \tag{2.16}$$

As a result of (2.13) above, the formula of coefficient alpha can be written as

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum\sigma_i^2}{\sum\sigma_i^2 + 2\left(\sum\sigma_{ij}\right)}\right]. \tag{2.17}$$

From this it is clear that the numerator and denominator of the ratio in the last factor will differ only when items comprising the total score are correlated. In the extreme case, when the correlation between all possible pairs of items is zero the total variance will equal the sum of the variances of items. Under such circumstances, the ratio of the sum of variances of the items to the total variance will equal 1 and the reliability coefficient will be 0. This is because the absence of correlations among the items means that they have nothing in common, which is a paradox to the concept of internal consistency reliability, namely that the items measure the same thing (Pedhazur & Schmelkin, 1991). It can be seen that according to (2.17) α increases if k (and therefore the number of covariances) gets larger, and thus researchers must not be mislead by a measure's internal consistency if it consists of a unrealistically large number of items.

In the case of dichotomously scored items, the Kuder-Richardson formula 20 for determining internal consistency is

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum p_i q_i}{\sigma_x^2}\right] \tag{2.18}$$

where p_i and q_i are the proportions in the sample of correct and incorrect answers for item *i*. According to Anastasi and Urbina (1997), Nunnally and Bernstein (1994) and Pedhazur and Schmelkin (1991) it is a special case for the general formula of α according to (2.12). It can be shown mathematically that the Kuder-Richardson formula 20 and Cronbach's alpha are both the mean of all split-half coefficients resulting from different splittings of a measure (Anastasi & Urbina, 1997).

2.4. Validity

2.4.1. Introduction

The validity of a test concerns what the test measures and how well it does so (Anastasi & Urbina, 1997). According to Aiken and Groth-Marnat (2006) a shortcoming of this definition is the implication that a test has only one validity. They say a test may have many different validities, depending on the specific purposes for which it was designed, the target sample, the conditions under which it is administered, and the method of determining validity. Pedhazur and Schmelkin (1991), however, took a strong stand against the notion of validity types, although they admit that it is convenient for organisation and discussion purposes. According to them the different facets of validity are not mutually exclusive and exhaustive and therefore there are not different types of validity. The latter perspective will be followed. Also keep in mind that while reliability is influenced only by unsystematic errors of measurement, the validity of a test is, however, affected by unsystematic as well as systematic (constant) errors. For this reason, a test may be reliable without being valid, but it cannot be valid without being reliable. The different facets of validity are discussed in 2.4.2.

2.4.2. Facets of Validity

A construct which is synonymous with a concept or a theoretical construction, aimed at organising and making sense out of our environment. The main purpose is to use observed variables to describe a construct or concept which is an unobservable variable like intelligence or anxiety (Pedhazur & Schmelkin, 1991).

Construct validity is the extent to which a test measures a theoretical concept or trait, such as a personality characteristic like intelligence. Construct validity can include measures of criterion-related validation, convergent validation, and content validation.

2.4.2.1. Criterion-related Validation

Criterion-related validation has two different facets, namely concurrent validation and predictive validation. Both are based on correlation (Aiken & Groth-Marnat, 2006; Anastasi & Urbina, 1997)

If a test is said to measure intelligence it must be shown that scores on the test are highly correlated with performance on an established test of intelligence (the standard or criterion for intelligence) (Aiken & Groth-Marnat, 2006; Anastasi & Urbina, 1997). In establishing concurrent validity, researchers administer the test to a group of participants and the scores

are correlated with a criterion measure (which is available for the participants) that reflects the variable being tested. For example, if the SAT 78 and GSAT had both been administered to a group of respondents, IQ scores of the SAT 78 are supposed to correlate highly with the scores of the subtest *Total* of the GSAT for both of the measures to have convergent validity, because they are both measures of intelligence.

Predictive validity is a facet of criterion-related validation where the criterion measures are obtained in the future, usually months or years after test scores are obtained (Nunnally & Bernstein, 1994). For example, the subtest *Total* of the GSAT will have predictive validity if high scores on this subtest go hand in hand with high scores for mathematics while low scores on this subtest go hand in hand with low scores for mathematics (3).

2.4.2.2. Convergent and Discriminant Validation

A construct-validated instrument should have high correlations with measures or methods of measuring the same construct (convergent validity), but low correlations with measures of different constructs (discriminant validity) (Anastasi & Urbina, 1997).

Convergent validation is a form of construct validity that refers to the degree that the actual test results are corresponding to the expected results. For example, if there was a test that measured *Numerical Ability*, and if it was high in convergent validity, it would be expected that individuals in risk management (namely the risk managers) of companies to score higher on the test than regular employees.

Discriminant validity is actually the complement of convergent validation. When a trait does not highly correlate with another trait that measures an unrelated concept it is said that the test has discriminant validity. For example, it would not be expected that leadership skills will highly correlate with shyness.

2.4.2.3. Content Validation

Content validation must not be confused with the term face validity which is a non-scientific judgment as to how well a test may superficially look to those who use it. It is, however, necessary for a test to have face validity because without it, cooperation and motivation as well as user and public acceptance will be problems (Linn, 1989).

Content validation addresses two questions:

1. Does the test cover the content of interest? For example, are the items on an achievement test for mathematics based on mathematical concepts?
2. Is the test appropriate for your participants? For example, are the items geared toward university mathematics majors?

Evaluating content validity is done in one of two ways, namely subjectively or empirically. Subjective methods involve asking experts to judge the relevance of the test items regarding the subject area being assessed. Empirical methods, such as principal components analysis and factor analysis, identify the underlying structure of the test items.

2.5. Psychometric Tests used in this Study

In the literature concerning standardised measuring instruments a lot of different terms are used for the questions and the linear combinations of the questions in the instrument. For the purpose of this study the different questions will be called items and the linear combinations of items or questions, subtests.

2.5.1. Senior Aptitude Test 78 (SAT 78)

2.5.1.1. Introduction

The SAT 78 which is a South African test, is a different test than the Scholastic Assessment Test (SAT) which is a reasoning test and is a standardised test for college admissions in the United States. The last mentioned SAT is administered by the public College Board in the United States and is developed, published, and scored by the Educational Testing Service (Frey & Detterman, 2003). The SAT 78 refers to the South African test.

The SAT 78 is widely used in South Africa by psychometrists and educational psychologists to give guidance to learners when choosing a career. In 2004 a report about the test use patterns and needs of psychological assessment practitioners were released under the leadership of Cheryl Foxcroft (Foxcroft *et al*, 2004). The report .was based on the findings of a postal survey, focus group interviews and individual interviews. The participants were all registered psychology professionals. According to the postal survey the SAT 78 was used by 34.6% of all practitioners who took part in the survey. The SAT 78 appeared on the list of the top 10 most used tests of psychometrists as well as counselling psychologists in South Africa, but it was not on the list of the top 10 tests used by research psychologists. The SAT 78 is also used by organisations' human resources departments for selection and placement purposes in different industries (van der Merwe, 1999) and also appeared on the list of top 10

tests used by industrial psychologists. This test was standardised in 1978 (Fouche & Verwey, 1978). In none of the subsequent published studies using this test, for purposes of guidance, prediction of academic success or placement in industry, were the reliability and validity of the test or subtests evaluated for the sample under consideration. Foxcroft *et al.* (2004) reports that only 63% of practitioners who took part in the survey, using the SAT 78, had information about the reliability and validity of the SAT 78.

2.5.1.2. Rationale

The Senior Aptitude Tests were compiled for measuring a number of aptitudes of pupils in grades 10 - 12 (previously standards 8, 9 and 10), and of adults. According to the developers of the tests, aptitude can be regarded as the potential which a person has and which enables him/her to attain a specific level of ability with a given amount of training and/or practice. Aptitudes, together with other personality characteristics such as interest, attitude and motivation, as well as training and instruction, will determine the level of skill and proficiency which may be reached.

The tests can also be used for guidance and selection purposes. It has also been established that a fairly reliable estimated IQ can be obtained with the aid of SAT 78 scores for testees in the age category 14 to 18 years. This fact enhances the practical values of the test battery.

2.5.1.3. Subtests of the SAT 78

The SAT 78 has 12 subtests. In this study the first 10 subtests were used. The last two subtests, namely subtests 11 and subtests 12 measuring co-ordination and writing speed, respectively, were not used in this screening process. It seems that although these two subtests had predictive validity (Fouche & Verwey, 1978), they were not regarded as measuring constructs that were appropriate in the selection and guiding process. They are also not used when calculating the estimated IQ.

Test 1: Verbal Comprehension

The test measures mainly the ability normally measured by verbal subtests of general intelligence. It is therefore mainly a measure of the general mental factor, G, which can be defined as the general level of cognitive functioning.

Test 2: Calculations

This test measures numerical ability, which is the ability to work quickly and correctly with figures. The items of the test cover basic skills like addition, subtraction, multiplication and division.

Test 3: Disguised Words

Since an understanding of the meaning of words is required, this test mainly measures the so-called AF Factor, Associational Fluency. The test also contains a component of the Verbal Reasoning Factor.

Test 4: Comparison

The test mainly measures the Visual Perception Speed Factor, P, of which the most important characteristic is speed and accuracy of perception of differences between, and similarities of visual configurations.

Test 5: Pattern Completion

This test measures the General Reasoning Factor, R. Since there is no mention here of mathematical problems, this test clearly measures the Inductive Reasoning Factor, I.

Test 6: Figures Series

This test mainly measures the General Reasoning Factor, R, which is a component of the General Mental factor, G. It is significant in this connection that Tests 5 and 6 both appear in the formula for estimating the IQ.

Test 7: Spatial 2 D

This test measures the Visualization Factor, V_z, and General Reasoning Factor, R.

Test 8: Spatial 3 D

This test also measures the General Reasoning Factor, R, and the Visualization Factor, V_z

Test 9: Memory (Paragraph)

The test requires the ability to memorise meaningful material and it measures the Memory Factor, M. The factor can be defined as the basic ability to memorise and to remember, irrespective of the complexity of the material.

Test 10: Memory (Symbols)

The test requires the ability to memorise meaningless material associatively. This test also measures the Memory Factor, M.

2.5.1.4. Reliability of SAT 78

At the time when the SAT 78 was standardised, reliability coefficients of the subtests of the SAT 78 were calculated using the Kuder-Richardson formula 8 (K-R 8) for subtests 1 to 10.

The original reliability coefficients of the subtests for the sample of size 1453 on which the test was standardised in 1978 as well as the reliability coefficients for our sample are given in Table 2.1.

Table 2.1 Reliability coefficients of SAT 78 subtests in 1978

Subtest	*K-R 8 (1978)*[1]
Verbal Comprehension	0.717
Calculations	0.921
Disguised Words	0.788
Comparison	0.762
Pattern Completion	0.834
Figure Series	0.852
Spatial 2D	0.918
Spatial 3D	0.838
Memory (Paragraph)	0.762
Memory (Symbols)	0.836

[1] From Fouche & Verwey, 1978, n = 1 453

2.5.1.5. Construct Validity of the SAT 78

The SAT 78 was standardised in 1978 using the following procedures. Construct validity was calculated through exploratory factor analysis (Fouche & Verwey, 1978). As can be seen in Table 2.2 in 1978 exploratory factor analysis yielded four significant factors. These four factors formed the four constructs that were given names such as *Verbal Ability* by the designers of the SAT 78. The SAT 78 manual does not state what portion of variation in the data was explained by these constructs.

Table 2.2 Original constructs and factor loadings of the subtests of the SAT 78

Constructs	*Factor Loadings*			
	Afrikaans		English	
	Boys	Girls	Boys	Girls
Construct 1 (Verbal Ability)				
Verbal Comprehension	0.62	0.66	0.52	0.52
Disguised Words	0.64	0.66	0.44	0.47
Memory (Paragraph)	0.43	0.39	0.45	0.38
Construct 2 (Numerical Ability)				
Calculations	0.55	0.64	0.66	0.62
Comparison	0.19	0.50	0.25	0.59
Construct 3 (Visual-Spatial Reasoning)				
Pattern Completion	0.51	0.54	0.49	0.52
Figure Series	0.58	0.61	0.44	0.49
Spatial 2D	0.62	0.60	0.68	0.49
Spatial 3D	0.72	0.60	0.71	0.67
Construct 4 (Memory)				
Memory (Paragraph)	0.46	0.47	0.36	0.45
Memory (Symbols)	0.37	0.60	0.45	0.52

From Fouche & Verwey, 1978, n = 1 453

2.5.2. General Scholastic Aptitude Test (GSAT)

2.5.2.1. Introduction

The GSAT is used in South Africa by psychometrists and educational psychologists to give guidance to learners when choosing a career. It is also used by faculties at universities for selection purposes (Engelbrecht, 1999). The GSAT is also used by researchers for prediction of academic success and it was found that a positive relationship between GSAT scores and academic achievement exists (Venter, 1995). De Bruin (1997), using both the SAT 78 and GSAT, recommended the GSAT being used for selection purposes such as academic success. According to van Eeden, de Beer and Coetzee (2001) the *Verbal* subtest of the GSAT contributed more than other psychometric tests to the variance in academic achievement.

This test was standardised in 1991 (Claassen, de Beer, Hugo & Meyer, 1991). The GSAT was not on any of the lists of the top 10 most used tests by the different types of psychologists (Foxcroft *et al.* 2004). According to the same report, 18.0% of the respondents in the postal survey used the GSAT.

2.5.2.2. Rationale

The General Scholastic Aptitude Test (GSAT) is a test to determine academic intelligence and scholastic aptitude (Claassen *et al.*, 1991). It has 6 subtests and takes about 160 minutes to complete. To determine a person's intellectual aptitude in a much shorter time two other versions of this test were also standardised, namely the abbreviated GSAT which has 4 subtests and takes about 110 minutes to complete and the abbreviated speed loaded GSAT which has the same 4 subtests as the abbreviated version and takes only 66 minutes to complete. All these versions were standardised in 1991 (Claassen *et al.*, 1991) and correlate highly with each other on their specific subtests.

According to Spearman's two factor theory of intelligence, each item in a cognitive test measures a general factor ***g*** and a specific factor ***s*** which is unique for each specific item (Nunnally & Bernstein, 1994). The main aim of the GSAT, therefore, was to choose items which gave a general indication of a person's general intellectual functioning namely ***g*** (Claassen *et al.*, 1991).

2.5.2.3. Subtests of the GSAT

Test 1: Word Analogies

The assumption is made here that the ability to recognize the relationship between two words and to complete another word pair in analogy to this gives a good indication of verbal reasoning.

Test 2: Figure Series

This subtest is based on the assumption that the ability to determine the relationship between numbers of a figure series, to deduce the rule and then apply it in the completion of the figure series gives a good indication of nonverbal reasoning

Test 3: Verbal Reasoning

The subtest is based on the assumption that the ability to determine relationships, to form new concepts and to manipulate them in a logical way is a good indication of an aspect of verbal reasoning.

Test 4: Pattern Completion

The assumption is made here that the ability to observe patterns accurately and to be able to complete the pattern is a good indication of an aspect of nonverbal reasoning.

2.5.2.4. Constructs of the GSAT

Verbal

This score combines the scores on test 1 and test 3 to provide a measure for verbal reasoning.

Nonverbal

This score combines the scores on Test 2 and Test 4 to provide a measure for nonverbal reasoning.

Total

This score combines the verbal and nonverbal scores.

2.5.2.5. Reliability of the GSAT

When the GSAT was standardised reliability coefficients for the *Verbal*, *Nonverbal*, and *Total* constructs were calculated using the Kuder-Richardson formula 8 (K-R 8).

The reliability coefficients of constructs for the sample (of size 138) in 1991 are given in Table 2.3.

Table 2.3 Reliability coefficients of GSAT constructs on the 1991 sample

Construct	*K-R 8 (1991)*[1]
Verbal	**0.91**
Word Analogies	
Verbal Reasoning	
Nonverbal	**0.91**
Figure Series	
Pattern Completion	
Total	**0.95**

[1] From Claassen *et al.*, 1991.
Subtest reliability coefficients were not reported, n = 138

2.5.2.6. Construct Validity of GSAT

When the GSAT was standardised construct validity was determined by exploratory factor analyses (Claassen *et al.*, 1991).

The factor loadings of the full version GSAT on the only significant principal component factor, which forms the only construct, are given in Table 2.4 for the original 1991 sample (of size 786).

Table 2.4 Factor loadings of the original subtests of the full version GSAT on the first principal component (1991)

Subtests	*Factor Loading*
Word Analogies	0.83
Word Pairs	0.85
Verbal Reasoning	0.89
Figure Series	0.83
Pattern Completion	0.82
Figure Analogies	0.85

From Claassen *et al.,* 1991, n = 786

The factor explained 50% of the variation in the data.

2.5.3. Brown-Holtzman Survey of Study Habits and Attitude (SSHA)

2.5.3.1. Introduction

The SSHA test is designed to evaluate study habits and study attitudes (du Toit, 1974). Eiselen and Geyser (2003) used the SSHA and found that achievers of academic success are more diligent than students who are at risk to fail. De Vetta (1987) found that white females scored invariably higher than white males on all four subscales of the SSHA and argued that it could be the reason why females surpass male scholars academically, in spite of the fact that the two sexes do not differ in mean IQ. In one of the few studies where reliability coefficients were calculated on the subtest of the SSHA, very low coefficients on all the subtests were obtained (Penny, 1984). The SSHA does not appear on any of the lists of the top 10 most used tests by the different types of psychologists (Foxcroft *et al.,* 2004). According to the same report, only 15.2% of the respondents used the SSHA.

2.5.3.2. Rationale

This questionnaire was developed in the USA by Dr. W.F. Brown and Dr. W.H. Holtzman and has been adapted and standardised in South Africa by the Institute for Psychometric Research of the Human Sciences Research Council. Two forms of the questionnaire are available, viz. Form H for secondary school learners and Form C for tertiary students. In this study Form H has been used, because the testees were either matriculants or university entrants, that is, before formally entering the university's environment. It is mainly used for the evaluation of respondents' study methods, their motivation for studying as well as certain attitudes with regard to academic activities in their learning environment. Diagnostically, this questionnaire may also give an indication of learners' habits and attitudes relating to scholastic activities and academic problems. The SSHA may be administered individually or in a group. No time limit applies and respondents receive sufficient time in which to complete the questionnaire. The test has 7 subtests. Every subtest consists of 25 items (du Toit, 1974).

2.5.3.3. Subtests of the SSHA

Test 1: Delay avoidance (DA)

This test indicates to what extent the learner promptly completes his or her assignment, avoids delay and is not inclined to unnecessary waste of time.

Test 2: Work method (WM)

This test gives an indication of the learner's use of effective study methods, efficiency in doing assignments and the extent to which he or she sets about his or her academic work in the most effective way.

Test 3: Study habits (SH)

This test combines the scores on the DA and WM scales to provide a measure for academic behaviour.

Test 4: Teacher approval (TA)

This test provides a measure of the learner's attitude towards the educator's classroom behaviour and methods.

Test 5: Education acceptance (EA)

This test determines the extent of the learner's acceptance of educational ideals, objectives, practices and requirements.

Test 6: Study attitudes (SA)

This test combines the scores of TA and EA to provide a measure of the learner's confidence in scholastic aims

Test 7: Study orientations (SO)

This test is a combination of all the above-mentioned aspects and provides an overall measure of the learner's study habits and attitudes.

2.5.3.4. Reliability of the SSHA

The SSHA's reliability coefficients of the subtests of the SSHA were calculated in 1974 using the split-half method.

The reliability coefficients of subtests for the sample (of size 2 790) in 1974 are given in Table 2.5. The manual does not report split-half coefficients for *Study Habits*, *Study Attitude*, and *Study Orientation*.

Table 2.5 Reliability coefficients of the SSHA subtests

Subtest	*Spit-half Coefficient (1974)*[1]
Delay Avoidance	0.833
Work Methods	0.835
Study Habits	
Teacher Approval	0.873
Education Acceptance	0.805
Study Attitude	
Study Orientation	

[1] From du Toit, 1974, n = 2 790

2.5.3.5. Construct Validity of the SSHA

Unlike with the SAT 78 and GSAT it seems from the SSHA's manual that construct validity was not determined through exploratory factor analysis (du Toit, 1974).

2.5.4. Personal Home Social and Formal Relations Questionnaire (PHSF)

2.5.4.1. Introduction

The purpose of the PHSF Relations Questionnaire is to measure, by means of 11 subtests, the personal, home, social and formal relations of high school pupils, students and adults, in order to determine the level of adjustment. By using the PHSF, Botha (1989) found that test anxiety in females is negatively related to social adjustment, and thus that lack of social adjustment could have a negative effect on academic performance. Naude, van Aarde and Laubscher (1989) came to the conclusion that some of the adjustment components, like family influences, moral sense, and formal relations could be the reason for better academic performance by English speaking students than Afrikaans speaking students. The PHSF does not appear on any of the lists of the top 10 most used tests by the different types of psychologists (Foxcroft *et al.,* 2004). According to the same report, 17.1% of the respondents in the survey used the PHSF.

2.5.4.2. Rationale

The level of adjustment of a person, for each of the various components of adjustment, is determined by the frequency with which his/her responses, in relations within the self or with the environment, are mature or immature, efficient of inefficient.

This does not imply concern with the measurement of personality traits as such, but rather with the expression and dynamics of these traits in the person's striving for harmony within the self and between the self and the environment. The PHSF includes 11 subtests of adjustment, which are divided into four main adjustment areas. A Desirability Scale is also included.

2.5.4.3. Constructs and subtests of the PHSF

I. Personal Relations (P)

This refers to the intra-personal relations which are of primary importance in adjustment.

Test 1: Self-confidence

This test refers to the degree to which a person has confidence in his ability, real or fancied, to be successful.

Test 2: Self-esteem

This test in an indication of the inner appraisal in a person based on evaluation and acceptance of real or fancied personality characteristics, abilities and defects.

Test 3: Self-control

This test refers to the degree to which a person succeeds in controlling and channelling his emotions and needs in accordance with his principles and judgement.

Test 4: Nervousness

A high score on this test indicates an absence of symptoms of nervousness as expressed by anxious, purposeless, repetitive behaviour.

Test 5: Health

A high score on this test indicates an absence of preoccupation with the physical condition.

II. Home Relations (H)

This test refers to the relations experienced by the person as a dependant within the family and home environment.

Test 6: Family Influences

This test is a measure of the degree to which a person as a dependant in a home is influenced by factors such as his position in the family, family togetherness, relationship between the parents, and socio-economic conditions.

Tests 7: Personal Freedom

This test is a measure of the degree to which a person feels that he is not restricted by his parents.

III. Social Relations (S)

This test refers to the manner in which a person engages in harmonious and informal relations within the social environment.

Test 8: Sociability-G

The test measures the degree to which a person has a need for and spontaneously participates in social group interaction (extrovert) in comparison with the degree to which a person is averse to social group interaction (introvert).

Test 9: Sociability-S

This test is a measure of the degree to which a person has a need for sociable interaction with a specific person of the opposite sex.

Test 10: Moral Sense

This test is a measure of the degree to which a person feels that his or her behaviour corresponds to the accepted norms of society.

IV. Formal Relations (F)

This test is an indication of the degree to which a person at school, college, university or in his occupation is successful in his formal relations with fellow pupils/fellow students/colleagues, as well as with figures of authority and superiors in the learning situation/work.

Test 11: Formal Relations

This test refers to the relations occurring in formal situations in the school, college or university, or occupation.

Test 12: Desirability Scale

This test is a validity scale indicating the honesty with which the person answered the questionnaire. The questions are of such a nature that only exceptional people can justly give favourable answers.

2.5.4.4. Reliability of the PHSF

Reliability coefficients for PHSF subtests were calculated in 1983 using the split-half method. As said in this study, Cronbach alpha coefficients were calculated to determine internal reliability.

The reliability coefficients of subtests for the samples of boys (size 909) and girls (size 879) in 1983 are given in Table 2.6.

Table 2.6 Reliability coefficients of the PHSF subtests

Subtest	*Spit-half Coefficient (1983)*	
	Boys	**Girls**
Self-confidence	0.80	0.79
Self-esteem	0.75	0.74
Self-control	0.71	0.70
Nervousness	0.74	0.74
Health	0.80	0.85
Family Influences	0.85	0.88
Personal Freedom	0.87	0.89
Sociability-G	0.88	0.89
Sociability-S	0.91	0.89
Moral Sense	0.79	0.77
Formal Relations	0.83	0.80
Desirability Scale	0.75	0.78

From Fouche & Grobbelaar 1983, n (boys) = 909, n (girls) = 879

2.5.4.5. Construct Validity of the PHSF

The PHSF's construct validity was determined in 1983 through exploratory factor analysis. Eight constructs (factors) were retained.

The factor groupings with highest loadings from PHSF manual are reported below. The manual does neither report the factor loadings nor the sample size or the percentage variation explained by the eight constructs displayed in Table 2.7.

Table 2.7 Constructs and subtests of PHSF (1983)

Construct 1 (Nervousness)
Nervousness
Health
Self-esteem
Self-control
Construct 2 (Home Relations)
Family Influences
Personal Freedom
Construct 3 (Moral Sense)
Moral Sense
Desirability Scale
Formal Relations
Self-control
Construct 4 (Sociability)
Sociability-G
Sociability-S
Construct 5 (Self-confidence)
Self-confidence
Formal Relations
Self-esteem
Construct 6 (School Relations)
Formal Relations
Construct 7 (Self-esteem)
Self-esteem
Construct 8 (Personal Freedom)
Personal Freedom

From Fouche & Grobbelaar, 1983

2.5.5. 19 Field Interest Inventory (19 FII)

2.5.5.1. Introduction

The purpose of the 19 FII is to determine in what types of activities the testee is most interested (Fouche & Alberts, 1977). Interest is defined as a relatively constant, positive or negative directedness towards a specific activity and is based on the whole personality. According to Foxcroft *et al.* (2004) the 19 FII was used by 51.8% of all practitioners in the survey. The 19 FII appears on the list of the top 10 most used tests of psychometrists, research psychologists, counselling psychologists, clinical psychologists, industrial psychologists, and educational psychologists in South Africa. Foxcroft *et al.* (2004) reports that only 55.7% of practitioners using the 19 FII had information about the reliability and validity of the 19 FII. Despite the fact that this inventory is used frequently, especially by research psychologists, no published research articles using the 19 FII could be found.

2.5.5.2. Rationale

The questions in the inventory refer to the pursuit of activities which underlie a number of the most important broad occupational fields. The person's directedness in respect of a certain group of activities should provide an indication of his interest in the vocational field(s) of which these activities form the basis. The inventory has 19 subtests.

2.5.5.3. Subtests of the 19 FII

Test 1: Fine Arts (FA)

Fine Arts embraces interest in activities which have bearing on painting, sculpture and sketching and also on the design of advertisements and signboards (commercial art).

Test 2: Clerical (Cl)

Clerical includes interest in routine work usually performed by clerks.

Test 3: Social Work (SW)

Interest in the rendering of service to the needy in society is covered by Social Work.

Test 4: Nature (Na)

Nature mainly refers to interest in activities which are pursued outdoors and covers stock farming, cultivation of crops and forestry.

Test 5: Performing Arts (PA)

Performing Arts has a bearing on interest in music, singing, ballet, opera and operetta.

Test 6: Science (Sc)

Science covers interest in the physical and biological sciences.

Test 7: Historical (H)

Historical gives an indication of the person's interest in the classics and in events which took place in the past.

Test 8: Public Speaking (PS)

Public Speaking refers mainly to the delivering of speeches and appearances in public.

Test 9: Numerical (Nu)

Numerical measures the person's interest in the use of numbers and other mathematical systems for the execution of calculations.

Test 10: Sociability (So)

Sociability is directed towards interest in social intercourse. It includes the organisation of, as well as participation in, social functions.

Test 11: Creative Thought (CT)

Creative Thought gives an indication of the person's interest in the use of logical thought for the solution of problems and in the execution of creative work.

Test 12: Travel (Tr)

Travel measures the extent to which persons like to travel often.

Test 13: Practical-Female (PF)

Practical-Female refers to interest in housekeeping, the making of clothes and other domestic activities which are pursued in the home, especially by women.

Test 14: Law (Lw)

Law refers to the study, as well as the application of laws and legal principles.

Test 15: Sport (Sp)

Sport gives an indication of the extent to which a person displays an interest in outdoor types of sport.

Test 16: Language (L)

Language includes interest in the appreciation of literature and the practical use and analysis of language.

Test 17: Services (Se)

Service refers to the rendering of service to persons in society who are not needy, such as, for example by waiters, shop assistants and hairdressers.

Test 18: Practical-Male (PM)

Practical-Male covers the mechanical and technical field and includes interest in the handling of tools for the practical execution of a task.

Test 19: Business (B)

Business includes interest in all forms of trading with a view to the making of a profit.

There are two additional tests, namely:

Test 20: Work-Hobby (W/H)

From this aspect of interest, an indication can be obtained whether a person is work or hobbies orientated in this interest.

Test 21: Active-Passive (A/P)

From this aspect it can be determined whether a person is actively interested in the pursuit of activities or whether he merely wishes to participate passively in these activities as a spectator.

2.5.5.4. Reliability of the 19FII

Reliability coefficients for 19 FII subtests were calculated using the split-half method in 1977.

The reliability coefficients of subtests for the samples in 1977 of boys (size 408) and girls (size 495) are given in Table 2.8.

Table 2.8 Reliability coefficients of the 19 FII subtests

Subtest	*Spit-half Coefficient (1983)*	
	Boys	**Girls**
Fine Arts	0.97	0.97
Performing Arts	0.95	0.97
Language	0.94	0.95
Historical	0.94	0.94
Service	0.92	0.9
Social Work	0.96	0.96
Sociability	0.96	0.95
Public Speaking	0.97	0.97
Law	0.98	0.98
Creative Thought	0.96	0.95
Science	0.97	0.95
Practical-Male	0.98	0.97
Practical-Female	0.96	0.96
Numerical	0.97	0.97
Business	0.98	0.97
Clerical	0.96	0.97
Travel	0.92	0.93
Nature	0.97	0.97
Sport	0.95	0.96
Work-Hobby	0.81	0.75
Active-Passive	0.73	0.68

From Fouche & Alberts 1977, n (boys) = 408, n (girls) = 495

2.5.5.5. Construct Validity

No method was mentioned in the manual how construct validity was determined for the 19 FII in 1977.

Fifteen fields of interest were identified in 1969 by Alberts in his DPhil thesis (cited in Fouche & Alberts, 1977). Neither factor scores, nor the sample size were reported in the manual in 1977 by Fouche and Alberts (see Table 2.9).

Table 2.9 Interest factors mentioned in 19 FII manual

Interest Factors	*Boys*	*Girls*
Fine Arts	Fine Arts	Fine Arts Creative Thought
Performing Arts	Performing Arts	Performing Arts
Language	Language	Language Historical
Service	Service Clerical Public Speaking	Service Clerical Public Speaking
Social Work	Social Work	Social Work
Sociability	Sociability Business Travel Sport	Sociability Business Travel Sport Service
Manipulation of scientific principles	Science	Science
Influencing the ideas and thinking of others	Public Speaking Law Language	
Manipulation of own thoughts and ideas	Creative Thought Science Nature	
Manipulation of thoughts and ideas		Public Speaking Law Creative Thought Business Language
Manipulation of things	Practical Male	Practical Male Nature
Manipulation of figures	Nature	Nature
Nature	Nature	
Evasion of occupational responsibility	Active-Passive Work-Hobby	
Travel	Travel Historical	

From Fouche & Alberts, 1977

Chapter 3

3. Literature study: Statistical Procedures

In this chapter the literature concerning the statistical procedures used in this study is discussed. The mathematical background and development of exploratory factor analysis, chi-squared automatic interaction detection (CHAID), logistic regression, discriminant analysis and effect sizes are given.

3.1. Exploratory Factor Analysis

3.1.1. Introduction

In exploratory factor analysis the aim is to describe and summarize data by grouping together variables that are correlated. One of the essential purposes of exploratory factor analysis is to describe the covariance relationship among many variables in terms of a few underlying unobservable quantities called factors. The motivation in exploratory factor analysis is to place variables in separate groups by their inter-correlations. A set of variables that highly correlate among themselves but have relatively small correlations with another set of variables will be placed in different groups. It is conceivable then that each group of variables represent a single underlying construct or factor that is responsible for the observed correlations (Johnson & Wichern, 2002).

3.1.2. The Orthogonal Factor Model

Factor analysis can be considered as an extension of principal component analysis. Both can be viewed as attempts to approximate the covariance matrix with fewer variables.

If the observable random vector $\mathbf{X}$, with p components, has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the factor model postulates that $\mathbf{X}$ is linearly dependent upon a few unobservable random variables $F_1, F_2, ..., F_m$, called common factors, and p sources of variation $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$, called errors or, sometimes specific factors. In particular, the factor analysis model is

$$
\begin{aligned}
X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1 \\
X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2 \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \varepsilon_p
\end{aligned}
\tag{3.1}
$$

or, in matrix notation,

$$
\underset{(p\times 1)}{\mathbf{X}} - \underset{(p\times 1)}{\boldsymbol{\mu}} = \underset{(p\times m)}{\mathbf{L}}\,\underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}} \tag{3.2}
$$

The coefficient ℓ_{ij} is called the *loadings* of the ith variable on the jth factor, so the matrix $\mathbf{L}$ is the *matrix of factor loadings*. Note that the ith specific factor ε_i is associated only with the ith response X_i. The p deviations $X_1 - \mu_1,\ X_2 - \mu_2,\ ...,\ X_p - \mu_p$ are expressed in terms of $p + m$ random variables $F_1,\ F_2,\ ...,\ F_m,\ \varepsilon_1,\ \varepsilon_2,\ ...,\ \varepsilon_p$ which are unobservable.

With so many unknown quantities, a direct verification of the factor model from observations on $X_1,\ X_2,\ ...,\ X_p$ is not possible. Thus, additional assumptions about the random vectors $\mathbf{F}$ and $\boldsymbol{\varepsilon}$, must be made, namely that the unobservable random vectors $\mathbf{F}$ and $\boldsymbol{\varepsilon}$ satisfy the following conditions:

$\mathbf{F}$ and ε are independent
$E(\mathbf{F}) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{F}) = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix
$E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\psi}$, where $\boldsymbol{\psi}$ is a diagonal matrix

The orthogonal factor model with m common factors is then

$$
\underset{(p\times 1)}{\mathbf{X}} = \underset{(p\times 1)}{\boldsymbol{\mu}} + \underset{(p\times m)}{\mathbf{L}}\,\underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}} \tag{3.3}
$$

and it implies the following covariance structure for $\mathbf{X}$, namely that

a) $\mathrm{Cov}(\mathbf{X}) = \mathbf{L}\mathbf{L}' + \boldsymbol{\psi}$

or $\mathrm{Cov}(X_i, X_k) = \ell_{i1}\ell_{k1} + \cdots + \ell_{im}\ell_{km}$ (3.4)

b) $\mathrm{Cov}(\mathbf{X}, \mathbf{F}) = \mathbf{L}$

or $\mathrm{Var}(X_i) = \ell_{i1}^2 + \cdots + \ell_{im}^2 + \psi_i$ and

c) $\mathrm{Cov}(X_i, F_j) = \ell_{ij}$.

The assumption is that the model $\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon}$ is linear in the common factors (Johnson & Wichern, 2002).

That portion of the variance of the ith variable contributed by the m common factors is called the ith communality. The portion of $\mathrm{Var}(X_i) = \sigma_{ii}$ due to the specific factor is often called the uniqueness, or specific variance.

Let the ith communality be denoted by h_i^2, then from (3.4) follows that

$$\underbrace{\sigma_{ii}}_{Var(X_i)} = \underbrace{\ell_{i1}^2 + \ell_{12}^2 + \cdots + \ell_{im}^2}_{\text{communality}} + \underbrace{\psi_i}_{\text{specific variance}}$$

and with

$$h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2 \tag{3.5}$$

therefore

$$\sigma_{ii} = h_i^2 + \psi_i, \quad i = 1, 2, ..., p.$$

The jth communality is the sum of squares of the loadings of the ith variable on the m common factors.

When $m > 1$, $\mathbf{L}$ and $\boldsymbol{\psi}$ are not unique. This ambiguity provides the rationale for "factor rotation", since orthogonal matrices correspond to rotations (and reflections) of the coordinate system for $\mathbf{X}$ (Johnson & Wichern, 2002). There are a number of different factor rotation methods such as varimax, quartimax, equimax, promax, and direct oblimin. In this study varimax rotation has been used which is an orthogonal rotation method. The reason for using it is that it was the method used in the standardisation procedures of all the psychometric tests. The loading matrix can be rotated (multiplied by an orthogonal matrix), where the rotation is determined by some "ease-of-interpretations" criterion. Once the loadings and specific variances are obtained, factors are identified, and estimated values for the factors themselves (called factor scores) are frequently constructed.

3.1.2.1. Methods of estimation

Given observations $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ on p generally correlated variables, factor analysis seeks to answer the question: Does the factor model of (3.3), with a small number of factors, adequately represent the data? In essence it is an attempt to find a statistical model which verifies the covariance relationship in (3.4).

The sample covariance matrix $\mathbf{S}$ is an estimator of the unknown population covariance matrix $\boldsymbol{\Sigma}$. If the off-diagonal elements of $\mathbf{S}$ are small or those of the sample correlation matrix $\mathbf{R}$ essentially zero, the variables are not related, and a factor analysis will not prove useful. In these circumstances, the specific factors play the dominant role, where the major aim of factor analysis is to determine a few important common factors.

Thus, if $\boldsymbol{\Sigma}$ appears to deviate significantly from a diagonal matrix, then a factor model can be entertained, and the initial problem is one of estimating the factor loadings ℓ_{ij} and specific variances ψ_i. Two of the most popular methods of parameter estimation are the principal component method and the maximum likelihood method. The solution from either method can be rotated in order to simplify the interpretation of factors. In this study the principal component method has been used.

Spectral decomposition of $\boldsymbol{\Sigma}$ provides us with one factoring of the covariance matrix (Johnson & Wichern, 2002).

Let $\boldsymbol{\Sigma}$ have eigenvalue-eigenvector pairs $(\lambda_i, \mathbf{e}_i)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then

$$\begin{aligned}\boldsymbol{\Sigma} &= \lambda_1 \mathbf{e}_1 \mathbf{e}_1' + \lambda_2 \mathbf{e}_2 \mathbf{e}_2' + \cdots + \lambda_p \mathbf{e}_p \mathbf{e}_p' \\ &= \left[\sqrt{\lambda_1}\mathbf{e}_1 \vdots \sqrt{\lambda_2}\mathbf{e}_2 \vdots \cdots \vdots \sqrt{\lambda_p}\mathbf{e}_p \right] \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1' \\ \hdashline \sqrt{\lambda_2}\mathbf{e}_2' \\ \hdashline \vdots \\ \hdashline \sqrt{\lambda_p}\mathbf{e}_p' \end{bmatrix}\end{aligned} \tag{3.6}$$

This fits the prescribed covariance structure for the factor analysis model having as many factors as variables $(m = p)$ and specific variances $\psi_i = 0$ for all i the loading matrix for jth column given by $\sqrt{\lambda_j}\mathbf{e}_j$. That is, it can be written

$$\underset{(p\times p)}{\boldsymbol{\Sigma}} = \underset{(p\times p)}{\mathbf{L}}\,\underset{(p\times p)}{\mathbf{L}'} + \underset{(p\times p)}{\mathbf{0}} = \mathbf{L}\mathbf{L}' \tag{3.7}$$

Apart from the scale factor $\sqrt{\lambda_j}$, the factor loadings on the jth factor are the principal component of the sample (that is the jth principal component is $\mathbf{e}_j'\mathbf{X}$, where the random vector $\boldsymbol{X}' = \left[X_i, X_2, \ldots, X_p \right]$ have the covariance matrix $\boldsymbol{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$) (Johnson & Wichern 2002).

Although the factor analysis representation of $\boldsymbol{\Sigma}$ in (3.7) is exact, it is not particularly useful: It employs as many common factors as there are variables and does not allow for any variation in the specific factors $\boldsymbol{\varepsilon}$ in (3.3). It is preferred that the models explain the covariance structure in terms of just a few common factors. One approach, when the last $p - m$ eigenvalues are small, is to neglect the contribution of $\lambda_{m+1}\mathbf{e}_{m+1}\mathbf{e}'_{m+1} + \cdots + \lambda_p\mathbf{e}_p\mathbf{e}'_p$ to $\boldsymbol{\Sigma}$ in (3.6). Neglecting this contribution, the following approximation is obtained

$$\boldsymbol{\Sigma} \doteq \left[\sqrt{\lambda_1}\mathbf{e}_1 \vdots \sqrt{\lambda_2}\mathbf{e}_2 \vdots \cdots \vdots \sqrt{\lambda_m}\mathbf{e}_m \right] \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1' \\ \hdashline \sqrt{\lambda_2}\mathbf{e}_2' \\ \hdashline \vdots \\ \hdashline \sqrt{\lambda_m}\mathbf{e}_m' \end{bmatrix} = \underset{(p\times m)}{\mathbf{L}}\,\underset{(p\times m)}{\mathbf{L}'} \tag{3.8}$$

The approximate representation in (3.8) assumes that the specific factors $\boldsymbol{\varepsilon}$ in (3.3) are of minor importance and can also be ignored in the factoring of $\boldsymbol{\Sigma}$. If specific factors are included in the model, their variances may be taken to be the diagonal elements of $\boldsymbol{\Sigma} - \mathbf{L}\mathbf{L}'$, where $\mathbf{L}\mathbf{L}'$ are defined in (3.8).

Allowing for specific factors, the approximation becomes

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\psi}$$

$$= \left[\sqrt{\lambda_1}\mathbf{e}_1 \vdots \sqrt{\lambda_2}\mathbf{e}_2 \vdots \cdots \vdots \sqrt{\lambda_m}\mathbf{e}_m\right] \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1' \\ \hdashline \sqrt{\lambda_2}\mathbf{e}_2' \\ \hdashline \vdots \\ \hdashline \sqrt{\lambda_m}\mathbf{e}_m' \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix} \tag{3.9}$$

where $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} \ell_{ij}$ for $i = 1, 2, \ldots, p$.

The representation in (3.8), when applied to the sample covariance matrix $\mathbf{S}$ or the sample correlation matrix $\mathbf{R}$, is known as the principal component solution (Johnson & Wichern, 2002). The name follows from the fact the factor loadings are the scaled coefficients of the first few sample principal components.

3.1.2.2. Principal Component Solution of the Factor Model

The principal component factor analysis of the sample covariance matrix $\mathbf{S}$ is specified in terms of its eigenvalues-eigenvector pairs $\left(\hat{\lambda}_1, \hat{\mathbf{e}}_1\right), \left(\hat{\lambda}_2, \hat{\mathbf{e}}_2\right), \ldots, \left(\hat{\lambda}_p, \hat{\mathbf{e}}_p\right)$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$. Let $m < p$ be the number of common factors. Then the matrix of estimated factor loadings $\{\tilde{\ell}_{ij}\}$ is given by

$$\tilde{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1 \vdots \sqrt{\hat{\lambda}_2}\hat{\mathbf{e}}_2 \vdots \cdots \vdots \sqrt{\hat{\lambda}_m}\hat{\mathbf{e}}_m\right] \tag{3.10}$$

The estimated specific variances are provided by the diagonal elements of the matrix $\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}'$, so

$$\tilde{\boldsymbol{\psi}} = \begin{bmatrix} \tilde{\psi}_1 & 0 & \cdots & 0 \\ 0 & \tilde{\psi}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\psi}_p \end{bmatrix} \text{ with } \tilde{\psi} = s_{ii} - \sum_{j=1}^{m} \tilde{\ell}_{ij} \tag{3.11}$$

Communalities are estimated as

$$\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \hat{\ell}_{i2}^2 + \cdots + \hat{\ell}_{im}^2 \tag{3.12}$$

The principal component factor analysis of the sample correlation matrix is obtained by starting with $\mathbf{R}$ in place of $\mathbf{S}$.

For the principal component solution the estimated loadings for a given factor do not change as the number of factors is increased. For example, if $m = 1$, $\tilde{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1\right]$, and if $m = 2$, $\tilde{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1 \vdots \sqrt{\hat{\lambda}_2}\hat{\mathbf{e}}_2\right]$, where $\left(\hat{\lambda}_1, \hat{\mathbf{e}}_1\right)$ and $\left(\hat{\lambda}_2, \hat{\mathbf{e}}_2\right)$ are the first two eigenvalues-eigenvector pairs for $\mathbf{S}$ (or $\mathbf{R}$).

By the definition of $\tilde{\psi}$, the diagonal elements of $\mathbf{S}$ are equal to the diagonal elements of $\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\boldsymbol{\psi}}$. However, the off-diagonal elements of $\mathbf{S}$ are not usually reproduced by $\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\boldsymbol{\psi}}$.

If the number of common factors is not determined by a priori considerations, such as by theory or the work of other researchers, the choice of m can be based on the estimated eigenvalues in much the same manner as with principal components. Consider the residual matrix

$$\mathbf{S} - \left(\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\boldsymbol{\psi}}\right) \tag{3.13}$$

resulting from the approximation of $\mathbf{S}$ by the principal component solution. The diagonal elements are zero, and if the other elements are also small, it may subjectively be taken that the m factor model is appropriate. Analytically the sum of squared entries of

$$\left(\mathbf{S} - \left(\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\boldsymbol{\psi}}\right)\right) \le \hat{\lambda}_{m+1}^2 + \ldots + \hat{\lambda}_p^2 . \tag{3.14}$$

Consequently, a small value for the sum of the squares of the neglected eigenvalues implies a small value for the sum of the squared errors of approximation.

Ideally, the contributions of the first few factors to the sample variances of the variables should be large. The contribution to the sample variance s_{ii} from the first common factor is $\tilde{\ell}_{i1}^2$. The contribution to the total sample variance, $s_{11} + s_{22} + \cdots + s_{pp} = tr\left(\mathbf{S}\right)$, from the first common factor is then

$$\tilde{\ell}_{11}^2 + \tilde{\ell}_{21}^2 + \cdots + \tilde{\ell}_{p1}^2 = \left(\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1\right)'\left(\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1\right) = \hat{\lambda}_1 .$$

Since the eigenvector $\hat{\mathbf{e}}_1$ has unit length, and in general,

$$\begin{pmatrix}\text{Proportion of total}\\ \text{sample variance}\\ \text{due to } j\text{th factor}\end{pmatrix} = \begin{cases}\dfrac{\hat{\lambda}_j}{s_{11}+s_{22}+\cdots+s_{pp}} & \text{for a factor analysis of } \mathbf{S}\\ \dfrac{\hat{\lambda}_j}{p} & \text{for a factor analysis of } \mathbf{R}\end{cases}. \qquad (3.15)$$

Criterion (3.15) is frequently used as a heuristic for determining the appropriate number of common factors. The number of common factors retained in the model is increased until a suitable proportion of the total sample variance has been explained. Another criterion to determine the number of factors to be extracted is Cattell's scree test. Eigenvalues, obtained from the analysis, are plotted and an inflection point of the resulting curve (scree) is determined by visual inspection. The location of the inflection point indicates the number of factors to be extracted. The third procedure, and the one used in this study, is Kaiser's Criterion, stating that as many factors should be extracted as factors with eigenvalues greater than or equal to one. The rationale behind this criterion is that interpretation of proportions of variance, smaller than the variance contribution of a single variable, are of dubious value. Kaiser's criterion is the one most frequently used since it does not require visual inspection of eigenvalue plots and is easily computerised (Hair, Anderson, Tatham & Black, 1998).

3.1.2.3. Practical Problems of Factor Analysis

A few practical issues and conditions must be taken into consideration before a factor analysis may be done.

Sample Size

Correlation coefficients tend to be less reliable when the datasets are small. According to Tabachnick and Fidell (2001) datasets with 500 or more observations are very good and thus adequate to assure that the results of a factor analysis will not be negatively affected by too few observations.

Outliers among cases

Outliers in the data have more influence on the factor solution than other observations and therefore care must be taken to identify multivariate outlier cases. The $n \times n$ matrix $\mathbf{H = X(X'X)^{-1}X'}$ is called the hat matrix, where

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

The hat matrix plays an important role in identifying influential observations in multiple regression. The hat matrix diagonal is a standardised measure of the distance of the *i*th observation vector of independent variables from the centroid of the x-space. If the diagonal elements of the hat matrix are denoted by h_{ii} (also called leverage values) (Belsley, Kuh & Welsch, 1980) then the relationship between the Mahalanobis distance and h_{ii} is given by

$$h_{ii} = \frac{\text{Mahalanobis distance for the i}^{\text{th}}\text{ observation}}{N-1} + \frac{1}{N} \qquad (3.16)$$

where N is the number of observations (Tabachnick & Fidell, 2001).

The Mahalanobis distance for each observation has a chi-square distribution, where the number of degrees of freedom depends on the dimension of the observation. Each case in a dataset can be evaluated to see if it may be classified as an outlier. A very conservative criterion for a case being an outlier is that $p < 0.001$ for the chi-square test. The chi-square value at significance level 0.001, with degrees of freedom equal to the number of variables, is then the maximum of what the Mahalanobis distance may be before the case is classified as an outlier (Tabachnick & Fidell, 2001).

Kaiser's measure of sample adequacy (MSA)

To determine whether a factor analysis may be appropriate, Kaiser's measure of sample adequacy (MSA), which gives an indication of the inter correlations among variables, should be computed (Tabachnick & Fidell, 2001). This index ranges from 0 to 1, reaching 1 when each variable is perfectly predicted by the other variables.

The measure can be interpreted with the following guidelines:

≥ 0.80: meritorious
0.70: middling
0.60: mediocre
0 50: miserable
< 0.50: unacceptable (Hair *et al.*, 1998).

Skewness and Kurtosis

According to Schepers (2004) the absolute value of the skewness of a variable used in a principal component factor analysis must not exceed 2, because it is disruptive and can lead to misleading factor structures. Furthermore, variables with kurtosis indices above 7 must be avoided and not be included in a factor analysis.

Singularity

Problems with the correlation matrix in factor analysis occur when variables are too highly correlated. With singularity, the variables are redundant which means that one of the variables is a linear combination of two or more of the other variables (Tabachnick & Fidell, 2001). To rule out singularity a good knowledge of the variables and where they came from is important (Montgomery, Peck & Vining, 2001; Tabachnick & Fidell, 2001).

3.2. Model Fitting Techniques

The goal of any model-building technique used in statistics is to find the best fitting and most parsimonious, yet practically reasonable model to describe the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables (Hosmer & Lemeshow, 2000). These independent variables can also be called *covariates*. The most common example of modelling is the linear regression model where the outcome variable is assumed to be continuous. Three other modelling techniques that can be used when the outcome variable is categorised are chi-squared automatic interaction detection (CHAID), logistic regression, and predictive discriminant analysis.

3.2.1. CHAID

3.2.1.1. Introduction

Classification trees are used to predict the membership of a case in the classes of a categorical dependent variable on one or more predictor variables. A decision tree is a non-linear discrimination method. CHAID is one of the most widely used methods of decision trees (Hair *et al.*,1998).

CHAID is an off-shoot of Automatic Interaction Detection (AID) and was originally proposed by Kass in 1980. CHAID partitions the data into mutually exclusive, exhaustive subsets that optimally predict the dependent variable (Kass, 1980). It operates on a nominal scale dependent variable and maximizes the significance of a chi-square statistic at each partition. More than two partitions are possible. The CHAID algorithm cannot handle continuous

independent (predictor) variables, therefore all the continuous predictor variables must be categorized into categorical data (Hawkins, 1982).

3.2.1.2. Types of Variables

Monotonic predictors are those whose categories lie on an ordinal scale, which implies that only neighbouring categories may be grouped together.

Free predictors are those whose categories are nominal.

Floating predictors are those where the categories lie on an ordinal scale with the exception of a single category that either does not belong with the rest or whose position on the ordinal scale is unknown.

3.2.1.3. Analysis Method

CHAID proceeds in steps. The first step is to create categorical predictors out of the continuous predictors, by dividing the continuous variables into a number of categories with the same amount of observations. Categorical predictors have already their specific categories.

The second step is searching through the predictors to find for each predictor the categories that are least significant with respect to the dependent variable. A chi-square test is then done. If the respective test for a given pair of predictors is not statistically significant according to a fixed alpha value, it will merge the predictor categories and the step will be repeated. On the other hand, if the test is significant a Bonferroni adjusted p-value for the set of categories will be computed (Kass, 1980).

The third step is to choose the predictor variable with the most statistically significant split. If the smallest p-value is greater than the fixed alpha to split value no further splits will be performed. The respective node is then a terminal node. If the p-value is smaller than this alpha value the process will continue until no further splits can be performed. Note that it is crucial that there are enough observations to ensure the validity of the chi-square test (Kass, 1980).

CHAID assumes that the effect of a variable in the subset is unrelated to the effect of the variable in other subsets. It naturally deals with interactions between the independent

variables that are directly available from an examination of the tree. The final nodes identify subgroups defined by different sets of independent variables.

3.2.1.4. Sample Size

Useful sample sizes for CHAID depend on the number of predictors, their types, the number of categories within each predictor, and how deep down the tree one would wish to go. Usually CHAID runs on thousands of observations. If there are only a few predictors with few categories, then a few hundred observations would suffice (Kass, 2008). The sample sizes in this study are small and therefore CHAID was only used for exploratory purposes.

3.2.1.5. Cross Validation

Cross validation involves splitting the sample into a number of smaller subsamples. Trees are than generated, excluding the data from each subsample in turn. For each tree, misclassification risk is estimated by applying the tree to the subsample excluded in generating it. The cross validated risk estimate for the overall tree is calculated as the average of the risks for all of these trees (SPSS Inc., 2007).

3.2.2. Logistic Regression

3.2.2.1. Introduction

Logistic regression is widely used in retention studies in higher education. According to Peng, So, Stage and John (2002b) a total of 52 articles published in three leading higher education research journals were identified as using logistic regression. These articles were published from 1988 to 1999 in Research in Higher Education, The Review of Higher Education, and The Journal of Higher Education. These three journals were cited by Silverman (1985) and Budd (1988) as core journals in higher education that publish research on a broad range of issues in this field (Budd, 1988). An in-depth search delivered no abstracts of articles in which logistic regression was used to predict academic success in any of these three journals after 2002. However, Houglum, Aparasu and Delfinis (2005) reported on predicting academic success in a Pharmacy Professional Programme using logistic regression. This article appeared in the American Journal of Pharmaceutical Education. Some research published in AIR Professional File, on retention studies, using *inter alia* logistic regression models, Cox regression models, and k-means cluster analysis, has been published after 2002, for example, Luo and Jamieson-Drake (2005) and Chen (2005).

In the case of categorical outcome variables, which are often used in retention studies, the linear regression model is inadequate. To overcome the limitations of least squares regression in handling categorical variables, a number of alternative statistical techniques have been used, for example logistic regression, discriminant analysis, Chi-squared Automatic Interaction Detector (CHAID), log-linear models, neural networks, and K-means cluster analysis.

Advantages of logistic regression are among others the following:

(a) It can accept both continuous and dichotomous predictors.
(b) It is not constrained by normality or equal variance/covariance assumptions for the residuals.
(c) It is related to the discriminant function analysis through Bayes theorem (Flury, 1997).

Furthermore, in terms of classification and prediction, logistic regression has been shown to produce fairly accurate results (Fan & Wang, 1999; Lei & Koehly, 2000). Therefore, researchers in higher education have recognised logistic regression as a viable alternative to predictive discriminant analysis and other techniques for analysing categorical outcome variables (Peng *et al.*, 2002b).

Although logistic regression is a relatively simple statistical procedure the mathematical development is quite complex. The reason for discussing the univariate (one predictor) as well as the multivariate case (more than one predictor) is that it is easier to use examples where there is only one predictor variable.

3.2.2.2. The Univariate Case

In a logistic regression model the outcome variable (Y) is binary or dichotomous (Hosmer & Lemeshow, 2000; Kutner, Nachtscheim, Neter & Li, 2005).

Let $\pi(x) = E(Y \mid x)$, where $E(Y \mid x)$ is the conditional mean of Y given the predictor x, the logistic regression model is then

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \tag{3.17}$$

The following logit transformation is made:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$$

then

$$g(x) = \beta_0 + \beta_1 x. \tag{3.18}$$

If the value of the outcome variable, given x, is denoted by $y = \pi(x) + \varepsilon$, the ε-term which is called the error, may assume two possible values. If $y = 1$ then $\varepsilon = 1 - \pi(x)$ with probability $\pi(x)$, and if $y = 0$ then $\varepsilon = -\pi(x)$ with probability $1 - \pi(x)$. Thus ε has a distribution with mean zero and variance equal to $\pi(x)\left[1 - \pi(x)\right]$ (Hosmer & Lemeshow, 2000).

Thus, when the outcome variable is dichotomous:

(1) The conditional mean of the regression $\pi(x) = E(Y \mid x)$ equation must be formulated to be bounded between 0 and 1 and (3.17) satisfies this constraint.

(2) The binomial distribution describes the distribution of errors, and will be the statistical distribution upon which the analysis is based.

The maximum likelihood method will be the approach followed to estimate the parameters in the logistic regression model. In a very general sense the method of maximum likelihood yields values for the unknown parameters which maximize the probability of obtaining the observed set of data. In order to apply this method a likelihood function must be constructed.

If Y is coded 0 or 1 then the expression for $\pi(x)$ given in equation (3.17) provides (for a value of $\boldsymbol{\beta} = (\beta_0, \beta_1)$, the vector of parameters) the conditional probability that Y is equal to 1 given x. This will be denoted as $P(Y = 1 \mid x)$. It then follows that the quantity $1 - \pi(x)$ gives the conditional probability that Y is equal to zero given x, $P(Y = 0 \mid x)$. Thus, for those pairs (x_i, y_i), where $y_i = 1$, the contribution to the likelihood function is $\pi(x_i)$, and for those pairs where $y_i = 0$, the contribution to the likelihood function is $1 - \pi(x_i)$, where the quantity $\pi(x_i)$ denotes the value of $\pi(x)$ computed at x_i. A convenient way to express the contribution to the likelihood function for the pair (x_i, y_i) is through the expression

$$\pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{1-y_i}. \tag{3.19}$$

Since the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in expression (3.19) as follows:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{1-y_i}. \tag{3.20}$$

The principle of maximum likelihood states that the estimate of $\boldsymbol{\beta}$ takes the value which maximizes the expression in equation (3.20). The log-likelihood is defined as

$$L(\boldsymbol{\beta}) = \ln\left[l(\boldsymbol{\beta})\right] = \sum_{i=1}^{n}\left\{y_i \ln\left[\pi(x_i)\right] + (1-y_i)\ln\left[1-\pi(x_i)\right]\right\}. \tag{3.21}$$

By partially differentiating $L(\boldsymbol{\beta})$ with respect to β_0 and β_1 respectively, the value of $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta})$ can be find by setting the resulting expressions equal to zero. The equations, known as the likelihood equations, are:

$$\sum\left[y_i - \pi(x_l)\right] = 0 \tag{3.22}$$

and

$$\sum x_i\left[y_i - \pi(x_i)\right] = 0. \tag{3.23}$$

In equations (3.22) and (3.23) the summation is over i varying from 1 to *n*.

For logistic regression the expressions in equations (3.22) and (3.23) are nonlinear in β_0 and β_1, and thus require iterative methods for their solution, and have been programmed into available logistic regression software (Hosmer & Lemeshow, 2000). The value of $\boldsymbol{\beta}$ given by the solutions to equations (3.22) and (3.23) will be denoted by $\hat{\boldsymbol{\beta}}$.

3.2.2.3. Testing for the Significance of the Coefficients

One approach to testing for the significance of the coefficient of a variable in any model relates to the following question: Does the model that includes the variable in question tell

more about the outcome (or response) variable than a model that does not include that variable? The question is asked in a relative sense. The comparison of observed to predicted values using the likelihood function is based on the following expression:

$$D = -2\ln\left[\frac{(\text{likelihood of the fitted model})}{(\text{likelihood of the saturated model})}\right], \tag{3.24}$$

where the predicted values of a saturated model are the observations themselves.

The quantity inside the square brackets in the expression in (3.24) is called the likelihood ratio. Using minus twice its log is necessary to obtain a quantity for which the asymptotic distribution is known and it can therefore be used for hypothesis testing purposes. Such a test is called the likelihood ratio test.

The statistic, *D*, in equation (3.24) is called the deviance and plays a central role in some approaches to assessing goodness-of-fit. Furthermore, if the values of the outcome variable are either 0 or 1, the likelihood of the saturated model is 1. Specifically, it follows from the definition of a saturated model that $\hat{\pi}_i = y_i$ and the likelihood is

$$l(\text{saturated model}) = \prod_{i=1}^{n} y_i^{y_i} \times (1-y_i)^{(1-y_i)} = 1.$$

Thus it follows from equation (3.24) that the deviance is

$$D = -2\ln(\text{likelihood of the fitted model}). \tag{3.25}$$

Using equation (3.21), equation (3.25) becomes

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right], \tag{3.26}$$

where $\hat{\pi}_i = \hat{\pi}(x_i)$, with the maximum likelihood estimators substituted for the parameters.

For purposes of assessing the significance of an independent variable a comparison of the value of D with and without the independent variable in the equation is made. The change in D due to the inclusion of the independent variable in the model is obtained as:

$$G = D(\text{model without the variable}) - D(\text{model with the variable}).$$

Because the likelihood of the saturated model is common to both values of D being differenced to compute G, it can be expressed as

$$G = -2\ln\left[\frac{(\text{likelihood without the variable})}{(\text{likelihood with the variable})}\right] \tag{3.27}$$

Under the hypothesis that β_1 is equal to zero, the statistic G follows a chi-square distribution with 1 degree of freedom, for n sufficiently large.

Two other similar, statistically equivalent tests have been suggested. These are the Wald test and the Score test. The assumptions needed for these tests are the same as those of the likelihood ratio test in equation (3.27).

The Wald test is obtained by comparing the ratio of the maximum likelihood estimate of the slope parameter, $\hat{\beta}_1$, to an estimate of its standard error. The resulting ratio, under the hypothesis that $\beta_1 = 0$, will asymptotically follow the standard normal distribution. The Wald test statistic for the logistic regression model is

$$W = \frac{\hat{\beta}_1}{\widehat{SE}\left(\hat{\beta}_1\right)}.$$

A test for the significance of $\hat{\beta}_1$ which does not require these computations is the Score test. The Score test is based on the distribution theory of the derivatives of the log likelihood. In the univariate case, this test is based on the conditional distribution of the derivative in equation (3.23), given the derivative in equation (3.22). The test statistic for the Score test (ST) is

$$ST = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y}) \sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

and it approximately follows the standard normal distribution.

3.2.2.4. The Multivariate Case

Consider a set of p independent variables denoted by the vector $x' = (x_1, x_2, \ldots, x_p)$. It will be assumed that each of these variables is at least on an interval scale. Let the conditional probability that the outcome is present be denoted by $y = 1$. The logit of the multiple logistic regression model is given by the equation

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{3.28}$$

in which case the logistic regression model is

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \tag{3.29}$$

3.2.2.5. Fitting the Multiple Logistic Regression Model

Assume a sample of n independent observations $(\mathbf{x}_i, y_i), i = 1, 2, \ldots, n$. As in the univariate case, fitting the model requires that estimates of the vector $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ are obtained. The method of estimation used in the multivariable case will be the same as in the univariate situation, namely maximum likelihood. The likelihood function is similar to that given in equation (3.20) with the only change being that $\pi(\mathbf{x})$ is now defined as in equation (3.29). There will be $p+1$ likelihood equations that are obtained by differentiating the log likelihood function with respect to the $p+1$ coefficients (Hosmer & Lemeshow, 2000).

The method of estimating the variances and covariances of the estimated coefficients follows from the theory of maximum likelihood estimation (Hosmer & Lemeshow, 2000). This theory states that the estimators are obtained from the matrix of second partial derivatives of the log likelihood function. These partial derivatives have the following general form:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1-\pi_i) \tag{3.30}$$

and

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1-\pi_i) \tag{3.31}$$

where π_i denotes $\pi(\mathbf{x}_i)$. Let the $(p+1)\times(p+1)$ matrix containing the negative of the terms given in equations (3.30) and (3.31) be denoted as $\mathbf{I}(\boldsymbol{\beta})$. This matrix is called the observed information matrix. The variances and covariances of the estimated coefficients are obtained from the inverse of this matrix which is denoted as $\text{Var}(\boldsymbol{\beta}) = \mathbf{I}^{-1}(\boldsymbol{\beta})$. Except in very special cases it is not possible to write an explicit expression for the elements in this matrix. Hence, the notation $\text{Var}(\boldsymbol{\beta}_j)$ will be used to denote the j^{th} diagonal element of this matrix, which is the variance of $\hat{\beta}_j$, and $\text{Cov}(\beta_j, \beta_l)$ to denote an arbitrary off-diagonal element, which is the covariance of $\hat{\beta}_j$ and $\hat{\beta}_l$. The estimators of the variances and covariances, which will be denoted by $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$, are obtained by evaluating $\text{Var}(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}$. $\widehat{\text{Var}}(\hat{\beta}_j)$ and $\widehat{\text{Cov}}(\hat{\beta}_j, \hat{\beta}_l), j, l = 0,1,2,\ldots,p$ will be used to denote the values in this matrix.

For the most part, it is only necessary to use the estimated standard errors of the estimated coefficients, which will be denoted as

$$\widehat{\text{SE}}(\hat{\beta}_j) = \left[\widehat{\text{Var}}(\hat{\beta}_j)\right]^{1/2} \tag{3.32}$$

for $j = 0,1,2,\ldots,p$. This notation will be used in developing methods for coefficient testing.

A formulation of the information matrix which will be useful when discussing model fitting and assessment of fit is $\hat{\mathbf{I}}(\hat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{V}\mathbf{X}$ where $\mathbf{X}$ is an n by $p+1$ matrix containing the data for each subject, and $\mathbf{V}$ is an n x n diagonal matrix with general element $\hat{\pi}_i(1-\hat{\pi}_i)$. That is, the matrix $\mathbf{X}$ is

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

and the matrix $\mathbf{V}$ is

$$\mathbf{V} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}.$$

3.2.2.6. Testing for the Significance of the Model

The likelihood ratio tests for overall significance of the p coefficients for the independent variables in the model are performed in the same manner as in the univariate case (Hosmer & Lemeshow, 2000). The test is based on the statistic G, with a chi-square distribution with p degrees of freedom. The univariate Wald test can be used to evaluate which variables are significant in the multivariate model.

The multivariate analogue of the Wald test is distributed as chi-square with p+1 degrees of freedom under the hypothesis that each of the p+1 coefficients is equal to zero. The multivariate analogue of the Score test for the significance of the model is based on the distribution of the p derivatives of $L(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. The computation of this test is of the same order of complication as the Wald test and hence there is no gain to use them instead of the likelihood ratio test (Hosmer & Lemeshow, 2000).

3.2.2.7. Interpretation of the Fitted Logistic Regression Model

In the case of one binary predictor $x = 0$ the difference in logits from (3.18) for $x = 1$ and $x = 0$ is:

$$g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1.$$

The first step in interpreting the effect of a covariate in a model is to express the desired logit difference in terms of the model. In this case the logit difference is equal to β_1. In order to interpret this result a measure of association termed the odds ratio is introduced and discussed.

The odds of the outcome being present among individuals with $x = 1$ is defined as $\pi(1)/[1-\pi(1)]$. Similarly, the odds of the outcome not being present among individuals with $x = 0$ is defined as $\pi(0)/[1-\pi(0)]$. The odds ratio, denoted OR, is defined as the ratio of the odds for $x = 1$ to the odds for $x = 0$, and is given by the equation

$$\text{OR} = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \tag{3.33}$$

and thus

$$\begin{aligned}
\text{OR} &= \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\Big/\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\Big/\left(\frac{1}{1+e^{\beta_0}}\right)} \\
&= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} \\
&= e^{(\beta_0+\beta_1)-\beta_0} \\
&= e^{\beta_1}.
\end{aligned}$$

Hence, for logistic regression with one dichotomous independent variable coded 1 and 0, the relationship between the odds ratio and the regression coefficient is

$$\text{OR} = e^{\beta_1}. \tag{3.34}$$

This fundamental relationship between the regression coefficient and the odds ratio is the reason why logistic regression has proven to be such a powerful analytic research procedure (Hosmer & Lemeshow, 2000; Kutner *et al.,* 2005). The relationship between the logistic regression coefficient and the odds ratio provides the basis for interpretation of all logistic regression results. This relationship can be extended to the multivariate case by keeping al the other *xs* constant.

When a logistic regression model contains a continuous independent variable, interpretation of the estimated coefficient depends on how it is entered into the model and the particular units of the variable. The primary difference is that a meaningful change must be defined for the continuous variable, namely *c*. The interpretation of the estimated coefficient for a continuous variable is then similar to that of nominal scale variables, given that there is a

linear relationship between the specific continuous variable and the logit. In the presence of interaction special care must be taken to interpret the odds ratio.

3.2.2.8. Variable Selection

The criteria for including a variable in a model may vary from one problem to another and from one scientific discipline to the next. The traditional approach to statistical model building involves seeking the most parsimonious model that will still explain the data. The rationale for minimizing the number of variables in the model is the fact that the resultant model is more likely to be numerically stable, and is more easily generalised. The more variables included in a model, the greater the estimated standard errors become, and the more dependent the model becomes on the observed data (Hosmer & Lemeshow, 2000). It is therefore essential to select variables carefully.

Stepwise Selection

Stepwise selection of variables is often used in logistic regression. This method is used especially in the case when there is a large number of independent variables available like in this study (Montgomery *et al.*, 2001).

Any stepwise procedure for selection or deletion of variables from a model is based on a statistical algorithm that checks for the importance of variables, and either includes or excludes them on the basis of a pre-determined decision rule. A measure of the statistical significance of the coefficient for the variable defines the "importance" of a variable. The statistic used depends on the assumptions of the model. In logistic regression the errors are assumed to follow a binomial distribution, and significance is assessed via the likelihood ratio chi-square test. The most important variable, in statistical terms, at any step in the procedure thus is the one that produces the greatest change in the log-likelihood relative to a model not containing the variable (i.e., the one that would result in the largest likelihood ratio statistic, G) (Hosmer & Lemeshow, 2000).

It is well known that p-values calculated in stepwise selection procedures are not p-values in the traditional hypothesis testing context. Instead, they should be thought of as indicators of relative importance among variables. The variables that had been selected should then be subjected to the more intensive analysis described in the previous section.

In **stepwise selection**, an attempt is made to remove any insignificant variables from the model before adding a significant variable to the model.

The **forward selection** technique starts with a default model and adds the most significant variables to the model according to specified criteria.

The **backward elimination** analysis starts with a model that contains all independent variables or covariates given in the model statement. The backward method eliminates the least significant variables (SAS Institute Inc., 2005b).

The **best subset of p predictor variables** is an alternative to stepwise selection of variables for a model. It and can be obtained by using SAS. The criterion used to determine the best subset of p predictor variables is based on the global score chi-square statistic. For two models A and B, each having the same number of explanatory variables, model A is considered to be better than model B if the value of the global score chi-square statistic C for A exceeds that for B (SAS Institute Inc., 2005b).

3.2.2.9. Assessing the Fit of the Model

When the efforts at the model building stage are at least preliminarily satisfactory, methods for assessing the fit of a logistic regression model can begin to be implemented. At this stage the model must contain those variables that should be in the model and the variables that have been entered in the correct functional form. The question now is how effectively this model describes the outcome variable. This is referred to as its goodness-of-fit (Hosmer & Lemeshow, 2000).

Suppose the observed sample values of the outcome variable in vector form are denoted as $\mathbf{y}$ where $\mathbf{y}' = (y_1, y_2, y_3, \ldots, y_n)$. The values predicted by the model, or fitted values, are denoted as $\hat{\mathbf{y}}$ where $\hat{\mathbf{y}}' = (\hat{y}_1, \hat{y}_2, \hat{y}_3, \ldots, \hat{y}_n)$. It is concluded that the model fits if summary measures of the distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$ are small and the contribution of each pair $(y_i, \hat{y}_i), i = 1,2,3,\ldots,n$ to these summary measures is unsystematic and relatively small to the error structure of the model (Hosmer & Lemeshow, 2000). However, summary statistics do not provide information about the individual predictors in the model.

Firstly, the effect the fitted model has on the degrees of freedom available for the assessment of model performance is considered. The term covariate pattern is used to describe a single set of values for the covariates in a model. For example, in a data set containing values of

gender, year group, qualification type, and verbal ability for each subject, the combination of these factors may result in as many different covariate patterns as there are subjects. On the other hand, if the model contains only gender and year group, each coded at two levels, there are only four possible covariate patterns. It is not necessary to be concerned about the number of covariate patterns during model development. The degrees of freedom for tests are based on the difference in the number of parameters in competing models, not on the number of covariate patterns. The number of covariate patterns may, however, be an issue when assessing the fit of a model (Hosmer & Lemeshow, 2000).

Goodness-of-fit is assessed over the constellation of fitted values determined by the covariates in the model, not the total collection of covariates. For instance, suppose that the fitted model contains p independent variables, $\mathbf{x}' = (x_1, x_2, x_3, \ldots, x_p)$, and let J denote the number of distinct values of $\mathbf{x}$ observed. If some subjects have the same value of $\mathbf{x}$ then $J < n$. The number of subjects are denoted with $\mathbf{x} = \mathbf{x}_j$ by $m_j, j = 1,2,3,\ldots,J$. It follows that $\sum m_j = n$. Let y_i denote the number of positive responses, $y = 1$, among the m_j subjects with $\mathbf{x} = \mathbf{x}_j$. It follows that $\sum y_j = n_1$, the total number of subjects with $y = 1$. The distribution of the goodness-of-fit statistics is obtained by letting n become large. If the number of covariate patterns also increases with n then each value of m_j tends to be small. Distributional results obtained under the condition that only n becomes large are said to be based on n-asymptotics. If $J < n$ is fixed and n is allowed to become large, then each value of m_j also tends to become large. Distributional results based on each m_j becoming large are said to be based on m-asymptotics. The difference between these asymptotics and the need to distinguish between them should become clearer as summary statistics are discussed in greater detail.

Initially, assume that $J \approx n$, as it can be expected whenever there is at least one continuous covariate in the model. This is the case most frequently encountered in practice (Hosmer & Lemeshow, 2000).

Pearson Chi-square Statistic and Deviance

In logistic regression there are several possible ways to measure the difference between the observed and fitted values. To emphasize the fact that the fitted values in logistic regression are calculated for each covariate pattern and depend on the estimated probability for that covariate pattern, the fitted value for the jth covariate pattern is denoted as $\hat{y}_j$ where

$$\hat{y}_j = m_j\hat{\pi}_j = m_j \frac{e^{\hat{g}(\mathbf{x}_j)}}{1+e^{\hat{g}(\mathbf{x}_j)}},$$

where $\hat{g}(\mathbf{x}_j)$ is the estimated logit.

The process is started by considering two measures of the difference between the observed and the fitted values: the Pearson residual and the deviance residual. For a particular covariate pattern the Pearson residual is defined as

$$r(y_j,\hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}. \tag{3.35}$$

The summary statistic based on these residuals is the Pearson chi-square statistic:

$$X^2 = \sum_{j=1}^{J} r(y_j,\hat{\pi}_j)^2 \tag{3.36}$$

The deviance residual is defined as

$$y_j - m_j\hat{\pi}_j, \tag{3.37}$$

where the sign, + or -, is the same as the sign of $y_j = 0$. For covariate patterns with $y_j = 0$ the deviance residual is

$$d(y_j,\hat{\pi}_j) = -\sqrt{2m_j\left|\ln(1-\hat{\pi}_j)\right.}$$

and the deviance residual when $y_j = m_j$, is

$$d\left(y_j,\hat{\pi}_j\right)=\sqrt{2m_j\left|\ln\left(\hat{\pi}_j\right)\right|}\,.$$

The summary statistic based on the deviance residuals is the deviance

$$D=\sum_{j=1}^{J}d\left(y_j,\hat{\pi}_j\right)^2 \tag{3.38}$$

In a setting where $J=n$, where each observation forms its own covariate pattern, this is the same quantity shown as in equation (3.37).

The distribution of the statistics X^2 and D under the assumption that the fitted model is correct in all aspects, is supposed to be chi-square with degrees of freedom equal to $J-(p+1)$. For the deviance this statement follows because D is the likelihood ratio test statistic of a saturated model with J parameters as opposed to the fitted model with $p+1$ parameters. Similar theory provides the null distribution of X^2. The problem is that when $J\approx n$, the distribution is obtained under the *m*-asymptotics, and hence the number of parameters increases at the same rate as the sample size. Thus, *p*-values calculated for these two statistics when $J\approx n$, using the $\chi^2(J-p-1)$ distribution, are incorrect.

One way to avoid these difficulties with the distributions of X^2 and D when $J\approx n$, is to group the data in such a way that *m*-asymptotics can be used. The rationale behind the various grouping strategies that have been proposed, is that X^2 can be thought of as the Pearson chi-square and D as the log-likelihood chi-square statistics that result from a $2\times J$ table. The rows of the table correspond to the two values of the outcome variable, i.e. for $y=1$ and $y=0$. J columns correspond to the J possible covariate patterns. The estimate of the expected value under the hypothesis that the logistic model in question is the correct model for the cell corresponding to the $y=1$ row and jth column is $m_j\hat{\pi}_j$. It follows that the estimate of the expected value for the cell corresponding to the $y=0$ row and jth column is $m_j\left(1-\hat{\pi}_j\right)$. The statistics X^2 and D are calculated in the usual manner from this table. The problem is that to expect these two statistics as coming from a chi-square distribution with $J-p-1$ degrees of freedom is not correct, because they are actually obtained from a $2\times J$ table.

The statistics that are arising from the $2\times J$ table cannot be expected to follow the $\chi^2(J-p-1)$ distribution. When chi-square tests are computed from a contingency table, the *p*-values are correct under the null hypothesis when the estimated expected values are sufficiently large in each cell. This condition holds under *m*-asymptotics. Although this oversimplifies the situation, it is essentially correct. In the $2\times J$ table the expected values are always quite small as the number of columns increases as *n* increases. To avoid this problem, the columns may be collapsed into a fixed number of groups, *g*, and the observed and expected frequencies are then calculated. By fixing the number of columns, the estimated expected frequencies become large as *n* becomes large. Thus, *m*-asymptotics hold (Hosmer & Lemeshow, 2000).

The Hosmer-Lemeshow Test

Hosmer and Lemeshow (2000) proposed grouping based on the values of the estimated probabilities. Suppose that $J=n$. In this case the *n* columns correspond to the *n* values of the estimated probabilities, with the first column corresponding to the smallest value, and the *n*th column to the largest value. Two grouping strategies were proposed as follows:

(1) collapse the table based on percentiles of the estimated probabilities, and
(2) collapse the table based on fixed values of the estimated probability.

With the first method, using $g=10$ groups results in the first group to contain the $n_1'=n/10$ subjects having the smallest estimated probabilities, and the last group to contain $n_{10}'=n/10$ subjects with the largest estimated probabilities. With the second method, using of $g=10$ groups results in cutpoints defined at the values $k/10, k=1,2,\ldots,9$, and the groups contain all subjects with estimated probabilities between adjacent cutpoints. For example, the first group contains all subjects whose estimated probability is less than or equal to 0.1, while the tenth group contains those subjects whose estimated probability is greater than 0.9. For the $y=1$ row, estimates of the expected values are obtained by summing the estimated probabilities over all subjects in a group. For the $y=0$ row, the estimated expected value is obtained by summing, over all subjects in the group, one minus the estimated probability. For either grouping strategy, the Hosmer-Lemeshow goodness-of-fit statistic, $\hat{C}$, is obtained by calculating the Pearson chi-square statistic from the $g\times 2$ table of observed and

estimated expected frequencies. Let c_k denotes the number of covariate patterns in the *k*th decile, and

$$o_k = \sum_{j=1}^{c_k} y_j$$

is the number of responses among the c_k covariate patterns, and

$$\overline{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n'_k}$$

is the average estimated probability. Then the following formula defines the calculation of $\hat{C}$:

$$\hat{C} = \sum_{k=1}^{g} \frac{\left(o_k - n'_k \overline{\pi}_k\right)^2}{n'_k \overline{\pi}_k \left(1 - \overline{\pi}_k\right)} \tag{3.39}$$

where n'_k is the total number of subjects in the *k*th group.

Hosmer and Lemeshow (2002) used an extensive set of simulations, when assuming $J = n$ and the fitted logistic regression model is the correct model, the distribution of the statistic $\hat{C}$ turned out to be well approximated by the chi-square distribution with $g-2$ degrees of freedom, $\chi^2(g-2)$.

Number of Predictors in the Model

Hosmer and Lemeshow (2000) suggest the following criterion for the number of predictors in the model: $\min(n_1, n_0)/10 \geq p+1$, where p is the number of predictors in the model, n_1 is the number of events and n_0 is the number of non-events.

Classification Tables

An intuitive way to summarize the results of a fitted logistic regression model is via a classification table. This table is the result of cross-classifying the outcome variable, *y*, with a dichotomous variable whose values are derived from the estimated logistic probabilities. These probabilities are derived from the leave-one-out principle; that is, dropping the data of one subject and re-estimating the parameter estimates. SAS uses a one-step approximation

to compute the parameter estimates. This option is only valid for binary response models (SAS Institute Inc. ,2005b).

To obtain the derived dichotomous variable, a cutpoint c must be defined and each estimated probability must be compared to c. If the estimated probability exceeds c, then the derived variable is considered to be equal to 1; otherwise it is equal to 0. The most commonly used value for c is 0.5. The appeal of this type of approach to model assessment comes from the close relationship of logistic regression to discriminant analysis when the distribution of the covariates is multivariate normal within the two outcome groups (Hosmer & Lemeshow, 2000).

In this approach, estimated probabilities are used to predict group membership. Presumably, if the model predicts group membership accurately according to some criterion, then this is thought to provide evidence that the model fits. Unfortunately, this is not necessarily the case. For example, it is easy to construct a situation where the logistic regression model is in fact the correct model and thus fits, but classification is poor. Accurate or inaccurate classification does not address the criteria for goodness-of-fit: that the distances between observed and expected values be unsystematic, and within the variation of the model. However, the classification table is useful.

Classification is sensitive to the relative sizes of the two component groups and always favours classification into the larger group, a fact that is also independent of the fit of the model. For practical purposes there is little difference between the values of $\hat{\pi} = 0.49$ and $\hat{\pi} = 0.53$, yet use of the 0.5 cutpoint would establish two individuals as markedly different.

Area under the Receiver Operating Characteristic (ROC) Curve

Sensitivity and specificity rely on a single cutpoint to classify a test result as positive or negative. A more complete description of classification accuracy is given by the area under the ROC curve. This curve, originating from signal detection theory, shows how the receiver operates the existence of a signal in the presence of noise. It plots the probability of detecting a true signal (sensitivity) and a false signal (1 – specificity) for an entire range of possible cutpoints.

The area under the ROC curve, which ranges from zero to one, provides a measure of the model's ability to discriminate between those subjects who experience the outcome of interest versus those who do not.

If the object was to choose an optimal cutpoint for the purposes of classification, a cutpoint may be selected that maximizes both sensitivity and specificity. A plot of sensitivity versus 1 – specificity over all possible cutpoints is called the ROC Curve and the area under the curve provides a measure of discrimination which is the likelihood that a subject who is a success will have a higher probability than a subject who is a failure.

As a general rule:

If the area under the ROC curve $= 0.5$:	this suggests no discrimination (i.e., the same as flipping a coin)
If $0.7 \leq area < 0.8$:	this is considered acceptable discrimination
If $0.8 \leq area < 0.9$:	this is considered excellent discrimination
If $area \geq 0.9$:	this is considered outstanding discrimination

An intuitive way of understanding the meaning of the area under the ROC curve is to create $n_1 x n_0$ pairs, where n_1 is the number of subject (respondents) with y = 1 and n_0 is the number of respondents with y = 0. By determining the proportion of time that the subject with y = 1 had the higher of the two probabilities it can be shown that this proportion is equal to the area under the ROC curve (Hosmer & Lemeshow, 2000).

It is, however, possible that a poorly fitting model may still have good discrimination (Hosmer & Lemeshow, 2000).

3.2.2.10. Logistic Regression Diagnostics

The summary statistics based on the Pearson chi-square residuals provide a single number that summarizes the agreement between observed and fitted values. A single number is used to summarize considerable information. Therefore, before concluding that the model "fits", it is important that other measures be examined to see if fit is supported over the entire set of covariate patterns (Hosmer & Lemeshow, 2000).

In logistic regression it is often best to rely on visual assessment, as the distribution of the diagnostics under the hypothesis that the model fits is known only in certain limited settings. It is impractical to consider all possible suggested plots in literature concerning influential data in logistic regression (Hosmer & Lemeshow, 2000).

The observations that are influential can be identified by inspecting certain plots because an influential point will clearly lie separate from the other data points in the Cartesian plane.

However according to Hosmer and Lemeshow (2000) the logistic regression model is remarkably flexible. Unless the dataset is of such nature that most of the probabilities are very small or very large, or in the case where the fit of the model is extremely poor, it is unlikely that any alternative model will provide a better fit. In this study no probabilities were very small or very large, thus regression diagnostics were not done.

3.2.3. Predictive Discriminant Analysisvery e

3.2.3.1. Introduction

There are two aspects of discriminant analysis (discriminant function analysis), namely descriptive discriminant analysis and predictive discriminant analysis (Huberty & Olejnik, 2006). In this study predictive discriminant analysis was used as a validation method to see how it performs in comparison with logistic regression to fit models to predict academic success. Two-group predictive discriminant analysis, with more than one predictor variable was used. However, the theory of predictive discriminant analysis is developed for more than two groups as well (Hawkins, 1982; McLachlan, 2004; Huberty & Olejnik, 2006).

Descriptive Discriminant Analysis (DDA)

Techniques of DDA are closely related to the study of effects determined by a MANOVA. As in other multivariate contexts linear combinations of the original multiple outcome variables are determined (this is where multivariate normality is assumed).

The primary questions addressed in DDA are:

1. How many constructs characterize group separation?
2. What constructs characterize group separation?

Predictive Discriminant Analysis (PDA)

PDA deals with prediction of group membership. In multiple regression analysis, a linear combination or the predictor variables are developed to predict the response (outcome) variable. In PDA a linear combination of predictors is also used. However, there are as many linear combinations as there are groups.

The primary questions addressed in a PDA are:

1. How accurately can group membership be predicted?
2. Is the resulting 'hit rate' better than that obtainable by chance?
3. If so, how much better?

These three questions are actually applicable in any type of predictive situation, that is, no matter whether logistic regression, linear regression or predictive discriminant analysis is used (Huberty & Olejnik, 2006).

Group Rule

A classification rule in PDA has three different forms. The first form is that of a combination of the predictor variables. The second form is that of an estimated probability of population membership and the third form is that of a distance between two points, generally the Mahalanobis distance under the normal model (Huberty & Olejnik, 2006). There are three types of distances, namely unit to unit (a unit in this study is a student), group to group (in this study successful group or failure group) and unit to group. The emphasis in PDA is on the last type (Huberty & Olejnik, 2006).

Maximum Likelihood, Posterior Probability, Prior Probability and Bayesian Probability

In predictive discriminant analysis a classification rule based on the **maximum likelihood principle** is used. That is, assigning a unit to the population in which its observation vector has the greatest likelihood of occurrence. Assuming that f is the multivariate density function for the J populations, $j = 0,1,2,\ldots,J$, the maximum-likelihood rule is: Assign unit u to population j if $f\left(x_u \middle| j\right) > f\left(x_u \middle| j'\right)$, $j' \neq j$.

The maximum-likelihood rule may also be stated in terms of **typicality probabilities** $P(\mathbf{x} \mid j)$, where $P(\mathbf{x} \mid j)$ is the probability that a unit has a profile close to $\mathbf{x}$, where $\mathbf{x}$ is the $p \times 1$ column vector of predictor variables, given that the unit is a member of population j.

The probability $P\left(j \middle| \mathbf{x}_u\right)$ is the **posterior probability** of membership in population j that is the probability of population membership of j after knowledge of $\mathbf{x}_u$, i.e. after the p, X values had been obtained. According to Huberty and Olejnik (2006) it is reasonable to assign unit u to population j if $P\left(\mathbf{x}_u \middle| j\right) > P\left(\mathbf{x}_u \middle| j'\right)$ for $j' \neq j$. Let π_j denote the proportion of units in

the universe that belongs to population j. Then if a unit comes from population j, before $\mathbf{x}$ is known the **prior probability** of membership in population j is π_j. Further, the product $\pi_j \cdot P(\mathbf{x}_u \mid j)$ is the joint probability that a unit belongs to population j and at the same time has a score vector close to $\mathbf{x}_u$. From these products and by using the Bayes rule.

$$P(j \mid \mathbf{x}_u) = \frac{\pi_j \cdot P(\mathbf{x}_u \mid j)}{\sum_{j'=1}^{J} \pi_{j'} \cdot P(\mathbf{x}_u \mid j')}. \tag{3.40}$$

The maximum (Bayesian) probability rule is then, assign unit u to population j if

$$P\left(j \middle| \mathbf{x}_u\right) > P\left(j' \middle| \mathbf{x}_u\right) \text{ for } j \neq j',$$

where $P\left(j \middle| \mathbf{x}_u\right)$ is defined as in (3.40).

By using the maximum likelihood rule and by minimising the total proportion of misclassification errors, the discriminant scores are obtained:

$$P(j \mid \mathbf{x}_u) = \frac{\pi_j \cdot f(\mathbf{x}_u \mid j)}{\sum_{j'=1}^{J} \pi_{j'} \cdot f(\mathbf{x}_u \mid j')}$$

for all the j populations. In this study's case j =1, 2 (Huberty & Olejnik, 2006).

To use these scores π_j and $f(\mathbf{x}_u \mid j)$ need to be estimated. Estimates of the 2 prior probabilities are sometimes (in this study's case) based on the sample sizes: $\widehat{\pi}_j = q_j = \frac{n_j}{N}$, with $N = \sum_j n_j$. These estimates are only appropriate if the sample sizes are in proportion to the population sizes. To estimate $f(\mathbf{x}_u \mid j)$ the assumption is made that the distribution of $\mathbf{x}$ in each of the j populations is the multivariate normal distribution.

If further the assumption is made that all the $\mathbf{x}$ s have equal variances, a linear rule must be used. In this case the linear classification function (LCF) is

$$L_{uj} = b_{1j}x_{1u} + b_{2j}x_{2u} + \cdots + b_{pj}x_{pu} + c_j, \tag{3.41}$$

where

$$c_j = -\frac{1}{2}\overline{x_j}\cdot\mathbf{S}_e\overline{x_j} + \ln q_j,$$

and $\mathbf{S}_e$ is the $p \times p$ pooled sample covariance matrix, and $\overline{x_j}$ is the mean vector of population j. Equation (3.41) can be written in many different forms.

If equal variances is not the case a quadratic rule must be used (McLachlan, 2004; Huberty & Olejnik, 2006). There are many different formulas for a quadratic classification function (QCF), for example

$$Q_{uj} = \ln q_j - \frac{1}{2}\ln\left|\mathbf{S}_j\right| - \frac{1}{2} + D_{uj}^2,$$

where $\mathbf{S}_j$ is the $p \times p$ covariance matrix for group j and $D_{uj}^2 = \left(\mathbf{x}_u - \overline{\mathbf{x}}_j\right)' \mathbf{S}_j^{-1}\left(\mathbf{x}_u - \overline{\mathbf{x}}_j\right)$.

Thus, the maximum-probability rule is: assign unit u to that population whose sample yields the largest QCF score (in the case of unequal covariance matrices) or the largest LCF score (in the case of equal covariance matrices) (Huberty & Olejnik, 2006).

3.2.3.2. Variable Selection

Just as in logistic regression, it is practical to use only certain predictor variables. **Forward selection** begins with no variables in the model. At each step, a variable is entered that contributes most to the discriminatory power of the model, as measured by Wilks' Lambda, the likelihood ratio criterion. When none of the unselected variables meets the entry criterion, the forward selection process stops (SAS Institute Inc. 2005b).

Backward elimination begins with all variables in the model. At each step, the variable that contributes least to the discriminatory power of the model as measured by Wilks' Lambda is removed. When all remaining variables meet the criterion to stay in the model, the backward elimination process stops (SAS Institute Inc. 2005b).

Stepwise selection begins, like forward selection, with no variables in the model. At each step, the model is examined. If the variable in the model that contributes least to the discriminatory power of the model as measured by Wilks' Lambda fails to meet the criterion to stay, then that variable is removed. Otherwise, the variable not in the model that contributes most to the discriminatory power of the model is entered. When all variables in the model meet the criterion to stay and none of the other variables meet the criterion to enter, the stepwise selection process stops (SAS Institute Inc. 2005b).

Cross-validation

If sample sizes are small the leave-one-out method instead of the hold-out method must be used. This method is based on leaving out the observation which has to be classified (Huberty & Olejnik, 2006).

Number of Predictors in the Model

To use the leave-one-out method for external validation a necessary condition is that $n_j > 3p$ where $n_j = \min(n_1, n_0)$ that is the number of observations of the group with the minimum number of observations where n_1 and n_0 are respectively the number of respondents in each of the two groups. p is the number of predictor variables in the model (Huberty & Olejnik, 2006).

Influential Observations

The study of influential observations in a predictive discriminant analysis is complicated and has been restricted to the two-group situation (Huberty & Olejnik, 2006). It was not used in this study.

3.2.4. Multicollinearity

With multicollinearity the variables are very highly correlated. A variance inflation factor is a measure of the degree to which and independent variable is correlated with the other independent variables in the model. Large variance inflation factors are indicators of multicollinearity and give an indication of which of the predictor variables are involved in the multicollinearity. Conditioning indexes can also be used to detect tightness or dependency of a variable on the others. Variance inflation factors of > 0.5 and conditioning index of > 0.3 is an indication of multicollinearity (Belsley *et al.*, 1980).

Multicollinearity is problematic for logistic regression analysis (Field, 2005), as well as predictive discriminant analysis (Naes & Mevil, 2001) because of the effect that it may have on the empirical inverse covariance matrix in both methods. As a result of multicollinearity this matrix becomes ill conditioned, which implies attempted division by very small numbers when the matrix inverse is calculated. Thus multicollinearity must be ruled out before starting the analyses.

3.3. Effect Sizes

3.3.1. Introduction

An advantage of drawing a random sample is it enables one to study the properties of a population with the time and money available. In such cases the statistical significance tests (e.g. t-tests) are used to show that the result (e.g. difference between two means) is significant. The *p*-value is a criterion of this, giving the probability that the obtained value (or more extreme) could be obtained under the assumption that the null hypothesis (e.g. no difference between the samples means) is true. A small *p*-value (e.g. smaller than 0.05) is considered as sufficient evidence that the result is statistically significant. Statistical significance does not necessarily imply that the result is important in practice as these tests have the tendency to yield *p*-values (indicating significance) getting smaller as the sizes of the data sets increase.

In many cases researchers are forced to consider their obtained results as a subsample of the target sample due to the weak response of the planned random sample. In other cases data are obtained from convenience sampling. These data should be considered as samples for which statistical inference and therefore p-values are not relevant. In addition to reporting descriptive statistics in these cases, effect sizes can be determined from which conclusions regarding practical significance can be drawn. Practical significance can be understood as a large enough difference to have an effect in practice.

Many different effect sizes exist (see Steyn, 1999) but in what follows, those relevant for this study will be discussed. When using effect sizes it should be kept in mind that the guidelines for different effect sizes are not rigid cutpoints.

3.3.2. Effect Size for Linear Relationships between two Continuous Variables

The Pearson moment correlation coefficient ρ_{xy} between the continuous variables x and y which is measured from population elements, is a measure of a linear relationship between x and y. Index ρ_{xy} is dimensionless with values between -1 and 1, where values 1 and -1 indicate a perfect linear relation and inverse linear relationship between x and y respectively and where $\rho_{xy} = 0$ means that there is no linear relationship between x and y. Because an effect size index must not depend on units, Cohen (1988) suggests that ρ_{xy} denoted only by ρ be used for an effect size index.

Guidelines for the Correlation Effect Size Index

If the correlations are used as effect sizes, the question is how large must it be to indicate an important relationship. Cohen (1988) suggests the following guidelines.

Small effect: $|\rho| = 0.1$
Medium effect: $|\rho| = 0.3$
Large effect: $|\rho| = 0.5$

Cohen (1988) motivated it as follows:

Small effect: $|\rho| = 0.1$, there is a small correlation which means that only 1% (i.e. $100 \times \rho^2 = 100 \times 0.1^2$) of the variation in y is explained by x.

Medium effect: $|\rho| = 0.3$, there is a medium correlation which means that about 10% of the variation in y is explained by x. This is a typical correlation in social sciences and such relationships can be observed by the naked eye.

Large effect: $|\rho| = 0.5$ means that 25% of the variation in y can be explained by x, which means that x and y are linearly related.

3.3.3. Effect Size for Goodness-of-fit Tests and of Independence based on Contingency Tables

Goodness-of-fit tests

In this case a single array of categories of sample frequencies is tested against a prescribed set of expected frequencies derived from the null hypothesis.

Tests of Independence

In this case sample observations are classified simultaneously by means of two different categorical variables to form a two-way frequency table. These frequencies are tested against the expected frequencies obtained when the null hypothesis of independence of the two variables holds.

In both cases an index which increases with the degree of discrepancy between the observed and expected frequencies must be found. For each category or combined category (cell) in a two-way table there are two proportions, one given by the null hypothesis and the other by what is observed. The effect size index w then measures the magnitude of the discrepancy between these paired proportions (Cohen, 1988):

$$w = \sqrt{\sum_{i-1}^{m} \frac{\left(P_{1i} - P_{0\mathrm{i}}\right)^2}{P_{0i}}} \tag{3.42}$$

where P_{0i} is the proportion in category i according to the null hypothesis;

P_{1i} the observed proportion of category i,

m is the number of categories or combined categories.

The value of w varies from 0 (when the paired Ps in all the cells are equal) to an upper limit of infinity.

In both of these cases it is important to know whether a relationship between two variables is practically significant. For a random sample, the statistically significance of such relationships is determined with chi-square tests, but the question is whether the relationship is large enough to be important.

In this case the effect size is given by $w = \sqrt{\frac{X^2}{n}}$, where X^2 is the usual chi-square statistic for the contingency table and n is the sample size (see Steyn, 1999 and 2002). In the special

case of a 2×2 table, the effect size(w) is given by the phi-coefficient, ϕ. Note that the effect size is again independent of sample size. Cohen (1988) gives the following guidelines for the interpretation of w:

small effect: $w = 0.1$

medium effect: $w = 0.3$

large effect: $w = 0.5$

An effect is practically significant if the effect size is large. To determine the practical effect of goodness-of-fit, a small effect size (defined as of no practical significance) would indicate that the fit is good (Ellis, 2002).

In the case of other tests involving the chi-square distribution, like the Wald test in logistic regression, the effect size of the regression coefficient of a dependent variable is to be determined, and this effect size is an indication of the importance of the variable, at that stage of the model building technique.

3.3.4. Effect Size of the Odds Ratio

The odds ratio has been defined in (3.33) as

$$\omega = \text{OR} = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}$$

The value of the OR lies between 0 and ∞, with value 1 if the two chance ratios (or odds) are equal. Thus, the OR can be evaluated in terms of how far it is from 1. The question, however, is how far from 1 will imply a practically significant effect.

According to Kline (2004) the OR can be transformed to a standardised difference, where

$$\begin{aligned}\delta_{OR} &= \frac{\ell n\left[\pi(1)/\left(1-\pi(1)\right)\right]-\ell n\left[\pi(0)/\left(1-\pi(0)\right)\right]}{1.81} \\ &= \frac{\text{logit}\left(\pi_1\right)-\text{logit}\left(\pi_0\right)}{1.81} \\ &= \frac{\ell n(\omega)}{1.81}\end{aligned} \tag{3.43}$$

According to Cohen (1988) the effect size guidelines for a standardised difference like (3.43) is

small effect $\delta_{OR} = 0.3$

medium effect: $\delta_{OR} = 0.5$

large effect : $\delta_{OR} = 0.8$

From (3.43) follows that $\omega = e^{1.81\delta_{OR}}$, (3.44)

therefore:

small effect: $\omega \approx 1.5$

medium effect: $\omega \approx 2.5$

large effect: $\omega \approx 4.25.$

By calculating the OR using ω the following guidelines for the odds ratio when the numerator odds are the larger than that of the denominator, so that OR ≥ 1 are suggested (Steyn, 2006):

significant (small): OR = 2.2

substantially significant (medium): OR = 2.5

highly significant (large): OR = 4.0

3.3.5. Effect Size Index for Improvement over Chance

In Table 3.1 a general classification table is given.

Table 3.1 General Classification Table

		Predicted Group		
		Success	Failure	Total
Actual Group	Success	a	b	a + b
	Failure	c	d	c + d
	Total	a + c	b + d	N

The number of correctly classified units according to Table 3.1, is a + d and is also the number of hits. The number of incorrect classified units is c + b. The observed hit rate H_o computed from Table 3.1 is

$$H_o = (a+d)/N.$$

This observed hit rate H_o that is determined by the leave-one-out method used for logistic regression and predictive discriminant analysis, gives an index of success of correct classification across populations or groups.

The question, however, is how much better than chance can the model predict group membership. In order to discuss the effect size index for improvement over chance, the proportional chance criterion is discussed and defined.

Proportional Chance Criterion

With J populations or groups of common size it could be expected to classify correctly about $\frac{1}{j}$ of the units by chance. Consider a situation where the J populations or groups are of varying sizes. Let q_j denotes the estimated prior probability of membership in group j then the chance frequency of hits for group j is $e_j = q_j n_j$, where n_j is the size of group j.

The overall or total group chance frequency of hits would be given as $e = \sum_{j=1}^{J} q_j n_j$, and the overall or expected chance hit rate is

$$H_e = \frac{e}{N} = \frac{1}{N}\sum_{j=1}^{J} q_j n_j. \tag{3.45}$$

The expression in (3.45) is the proportional chance criterion for the total group hit rate (Huberty & Olejnik, 2006).

By comparing the observed hit rate H_o with the expected hit rate H_e, H_o is corrected for coincidental correct classification of units. The chance error rate $1 - H_e$, in comparison with the observed error rate $1 - H_o$ are incorporated into the effect size index for improvement over chance.

$$I = \frac{(1 - H_e) - (1 - H_o)}{1 - H_e} = \frac{H_o - H_e}{1 - H_e} \tag{3.46}$$

(Steyn, 2006).

Guidelines for the Effect Size I

In the multivariate case (i.e. more than one predictor) guidelines for I are:

In the two group (population) case, when the covariance matrices of the two groups are equal,

$I < 0.15$: small effect

$0.2 < I < 0.3$: medium effect

$I > 0.35$: large effect

In the two group (population) case, when the two covariance matrices are not equal,

$I < 0.1$: small effect

$0.15 < I < 0.25$: medium effect

$I > 0.3$: large effect (Steyn, 2006).

Chapter 4

4. Methodology

In Chapters 2 and 3 attention was given to the literature on the psychometric tests, as well as the statistical procedures used in this study.

In this chapter the focus is on the methods used for the empirical investigation. Aspects that will be covered are the processing of the data, the study samples, the variables and the statistical techniques used.

4.1. Data Collection

The Department of Student Guidance of the North-West University's Potchefstroom Campus on the first Monday of every academic year conducts a battery of psychometric tests on their new first year students. Until 2002 this was compulsory for all students, but since 2003 it has been voluntary and students were encouraged to do these tests in order to ensure that they enrol for the appropriate courses that 'match' their abilities and fields of interests. However, for prospective Pharmacy students these tests were compulsory for selection purposes and were conducted in the September recess of their matric year (final school year), the year prior to their intended enrolment. The administration of the university also records the biographical data of each first year student as well as their matriculation results. In this study, these data of specific year groups were used in the prediction of academic success. From a statistical viewpoint, a student is considered an academic success if his or her degree was completed in the prescribed time, and as a failure if not.

4.2. Study Samples

As no random selection of students was made, the datasets can be regarded as available samples. Throughout this study effect size measures (Cohen, 1988; Ellis & Steyn, 2003; Huberty & Olejnik, 2006; Kline, 2004; Steyn, 1999, 2000, 2002, 2006) are used to determine practical significance since inferential statistics are not appropriate. The p-values obtained for the statistical procedures are reported, as if random sampling had been done.

4.3. Available Data

The data used in this study came from three main sources: the psychometric tests, the biographical and academic history data, and university graduation data. In the case of prospective Pharmacy students three further pre-registration tests were conducted on the same day when the psychometric tests were done and the marks obtained for these tests were also available: Mathematics, Physics, and Chemistry knowledge tests.

4.3.1. Psychometric Tests' Raw Data

The following psychometric tests have been used in this study: the Senior Aptitude Test 78 (SAT 78), the General Scholastic Aptitude Test (GSAT), the Brown-Holtzman Survey of Study Habits and Attitudes (SSHA), the Personal, Home, Social, and Formal Relations Questionnaire (PHSF), and the 19 Field Interest Inventory (19 FII).

All the psychometric tests were completed on answer sheets and the responses were entered by means of an optical (mark) reader. The student number of each respondent, appearing on every separate answer sheet, served throughout this study as an identification of the respondents.

The data were sent as text files from the optical reader and were then converted to SAS datasets (SAS Institute Inc., 2005a). SAS programmes to score these tests were written to obtain scores for all the subtests. Substantial data cleaning had to be done on the student numbers because respondents made mistakes by entering slightly different numbers on the different answer sheets or left out their number all together. As a result of this, problems occurred when trying to merge the data of all the psychometric tests by student number. The students' surname, initials, and gender also appeared on each test's answer sheet and by going back to all the student numbers which could not be merged, errors were corrected by comparing surname initials and gender to match to the correct student number. It was then entered manually on the dataset where the incorrect student number appeared and the merge could then be completed successfully.

4.3.2. Biographical and Academic History Data

The administration of the university also files the biographical data of each first year student as well as their matric results. The available academic history data are matric results. The matric results are available in terms of symbols of subjects (either higher grade or standard grade) obtained in their final matriculation examination. English (first language), Afrikaans

(first language), English (second language) and Afrikaans (second language) are all different subjects. The Pharmacy students in the Faculty of Health Sciences, the BCom students in the Faculty of Economic and Management Sciences, the BSc students in the Faculty of Natural Sciences, and the BA students in the Faculty of Arts were substantial groups and were considered suitable for this study. Therefore this study deals with these four groups separately for model fitting purposes. The number of available students in other degree courses was too small to be analysed by faculty or type of degree.

4.3.3. University Graduation Data

Graduation ceremonies at the Potchefstroom Campus of the North West University are held biannually, in March and in September. Most of the graduates at the March ceremonies obtained their bachelors degrees in the prescribed time. None of the students who received bachelor's degrees in September completed their degree in the prescribed time. This is true for both three and four year degrees. Thus, for the purpose of this study the March graduation data were used. There are also students who did not get the degree in the prescribed time at the March ceremonies but, by merging the student numbers, qualification code, and year of entrance, a dataset for the successful students was obtained. The rest of the students who had registered for the specific qualification code and in the specific year of entrance whose names and student numbers did not appear on the lists of graduandi are then for the purpose of this study termed as the failures.

4.3.4. Preparation of Datasets

4.3.4.1. Psychometric Tests' Data

Datasets for all the psychometric tests (SAT 78, GSAT, SSHA, PHSF and 19F11) have been created from the available psychometric tests' raw data to evaluate the reliability and validity of the tests and subtests of these psychometric tests used in this study. The procedure was done by combining different years' data for each separate test. Year groups 2003 to 2007 were used. The only necessary condition was to ensure that no student number appeared more than once.

4.3.4.2. Model Fitting Data

Eight datasets have been created to try to find models for predicting academic success. By merging the data coming from the three main sources, two datasets for BA, two for BSc, two for BCom, and two for BPharm degrees were created, using the classes of 2003 (group 1)

and 2004 (group 2). The data were verified so that the correct data for each student had been merged.

Table 4.1 gives the numbers of students per group.

Table 4.1 Study samples used for models

Degree	*Success n*	*Success%*	*Failure n*	*Failure%*	*Total n*	*Year*	*Group*
BPharm	100	76.34	31	23.66	131	2003	1
BPharm	108	79.41	28	20.59	136	2004	2
BPharm Total	208	77.90	59	22.10	267		
BA	24	32.00	51	68.00	75	2003	1
BA	40	39.22	62	60.78	102	2004	2
BA Total	64	36.16	113	63.84	177		
BSc	25	47.17	28	52.83	53	2003	1
BSc	34	35.05	63	64.95	97	2004	2
BSc Total	59	39.33	91	60.67	150		
BCom	96	59. 63	65	40.37	161	2003	1
BCom	114	50.89	110	49.11	224	2004	2
BCom Total	210	54.55	175	45.45	385		

4.4. Dependent Variable

In this study the outcome (dependent) variable (y) is status. In Table 4.2 a summary of the details of the dependent variable is given.

Table 4.2 Dependent variable

Variable	*Description*	*Codes/Values*	*Name*
1	Outcome variable (y)	success=1 failure=0	Status

4.5. Independent Variables

The choice of variables was mainly determined by the fact that the data of several psychometric tests were available. Biographical data such as gender and race were also used in this study. In the case of the Pharmacy students their marks obtained for the Mathematics test, Chemistry test, and Physics test at the time of selection, were also used in the model building process. Matric results, captured as symbols obtained per subject for each student, were transformed to an interval scale variable taking into account whether the subject was taken on higher or standard grade by using the information in Table 4.3. A score denoted by *Mscore* for each student (respondent) was then calculated by adding the weighted scores for each subject for each student. If a student had 6 matriculation subjects the *Mscore* was taken as the summation. If seven subjects were summed the *Mscore* was multiplied by $\frac{6}{7}$ and if eight subjects were summed the *Mscore* was multiplied by $\frac{6}{8}$.

In Table 4.3 the weights for conversion of symbols on higher or standard grade to numerical values are given (NWU, 2007).

Table 4.3 Conversion table of matric symbols to numerical weights

Symbol	*Higher Grade*	*Standard Grade*
A	6	5
B	5	4
C	4	3
D	3	2
E	2	1
F	1	0

From NWU (2007)

The information about the independent variables is given in Table 4.4:

Table 4.4 Independent variables

Variable	*Description*	*Codes/Values*	*Variable Type*
1	Race	1 = white 2 = coloured 3 = black 4 = Asian 5 = other	nominal
2	Gender	M = male V = female	nominal
3	Year Group	1 = year 2003 2 = year 2004	nominal
4	Mscore	10 - 36	interval scale
	SAT 78		
5	Verbal Comprehension	0 - 30	interval scale
6	Calculations	0 - 40	interval scale
7	Disguised Words	0 - 30	interval scale
8	Comparison	0 - 30	interval scale
9	Pattern Completion	0 - 30	interval scale
10	Figure Series	0 - 30	interval scale
11	Spatial 2D	0 - 30	interval scale
12	Spatial 3D	0 - 30	interval scale
13	Memory (Paragraph)	0 - 20	interval scale
14	Memory (Symbols)	0 - 30	interval scale
	SSHA		
15	Delay Avoidance	0 - 50	interval scale
16	Work Methods	0 - 50	interval scale
17	Teacher Approval	0 - 50	interval scale
18	Education Acceptance	0 - 50	interval scale
	PHFS		
19	Self-confidence	0 - 45	interval scale
20	Family Influences	0 - 45	interval scale
21	Moral Sense	0 - 45	interval scale
22	Health	0 - 45	interval scale
23	Self-esteem	0 - 45	interval scale
24	Sociability-S	0 - 45	interval scale
25	Formal Relations	0 - 45	interval scale
26	Nervousness	0 - 45	interval scale
27	Personal Freedom	0 - 45	interval scale

28	Self-control	0 - 45	interval scale
29	Sociability-G	0 - 45	interval scale
	19 FII		
30	Fine Arts	0 - 45	interval scale
31	Performing Arts	0 - 45	interval scale
32	Language	0 - 45	interval scale
33	Historical	0 - 45	interval scale
34	Service	0 - 45	interval scale
35	Social Work	0 - 45	interval scale
36	Sociability	0 - 45	interval scale
37	Public Speaking	0 - 45	interval scale
38	Law	0 - 45	interval scale
39	Creative Thought	0 - 45	interval scale
40	Science	0 - 45	interval scale
41	Practical-Male	0 - 45	interval scale
42	Practical-Female	0 - 45	interval scale
43	Numerical	0 - 45	interval scale
44	Business	0 - 45	interval scale
45	Clerical	0 - 45	interval scale
46	Travel	0 - 45	interval scale
47	Nature	0 - 45	interval scale
48	Sport	0 - 45	interval scale
	GSAT		interval scale
49	Word Analogies	0 - 25	interval scale
50	Verbal Reasoning	0 - 25	interval scale
51	Figure series	0 - 25	interval scale
52	Pattern Completion	0 - 25	interval scale
	Knowledge Tests		
53	Mathematics	1 - 100	interval scale
54	Physics	1 - 100	interval scale
55	Chemistry	1 - 100	interval scale

4.6. Processing the Data

The statistical computer package SAS (SAS Institute Inc., 2005a) was used throughout this study for creating permanent SAS datasets and calculating descriptive statistics. SAS was also used to calculate reliability of subtests and to perform exploratory factor analyses. SAS's procedures for logistic regression and discriminant analysis were used for model fitting purposes. In some instances SPSS (SPSS Inc., 2007) were also used. If the procedure was done in SPSS it will be specified, otherwise SAS was used as the computer package in this study.

4.7. Methods and Statistical Techniques

4.7.1. Methods Used to Address the First Research Question

As stated in Chapter 1, the first research question in this study is the following: Are the SAT 78, GSAT, SSHA, PHSF, and 19 FII reliable instruments for the study sample and, if so, how do the reliability of these instruments compare with the reliability at the time of their standardisation?

In this study Cronbach alpha coefficients were calculated to determine the internal reliability coefficients of the subtests of the different psychometric tests for the study samples. If the Cronbach alpha coefficient was 0.70 or more the specific subtest was considered reliable (Aiken & Groth-Marnat, 2006; Nunnally & Bernstein, 1994). The specific subtests may then be used as an independent variable in CHAID, logistic regression or discriminant analysis in order to address the third research question (Aiken & Groth-Marnat, 2006; Nunnally & Bernstein, 1994). In Chapter 5 a comparison between the reliability coefficients obtained for subtests in this study and those obtained when the test was standardised is made.

4.7.2. Methods used to Address the Second Research Question

As also stated in Chapter 1 the second research question in this study is the following: Are the SAT 78, GSAT, SSHA, PHSF, and 19 FII construct valid instruments for the study sample and, if so, how does the construct validity of these instruments compare with the construct validity at the time of standardisation?

Principal component exploratory factor analyses were done to determine construct validity for the different psychometric tests. The subtests of the specific psychometric tests were the variables used in the factor analysis. If more than one factor was extracted a varimax

rotation was done. In Chapter 3 conditions which had to be met for a factor analysis to be valid are discussed.

Singularity was ruled out (see Section 3.1.2.3). In this study, for example, IQ is a linear combination of some of the subtests of the SAT 78 and thus could have yielded redundant variables if it was used in the factor analysis, together with the individual subtests in the formula for IQ of the SAT 78. The subtests and not the constructs (that is a linear combination of subtests) of the specific psychometric tests were the variables used in the factor analysis. Care had been taken throughout that no redundant variables were used in the factor analysis.

Outlier cases in the data were detected by using leverage (hat) values (see Section 3.1.2.3). These values are related to the Mahalanobis distance discussed in Section 3.1.2.3. The cases that were identified as outliers had then been checked to assure that they were part of the valid observations and were kept in the data (Tabachnick & Fidell, 2001). A contingency table of race versus outliers/non-outliers was also computed to assure that race was not the variable responsible for outliers in the data. The phi coefficient, *w*, as discussed in Section 3.3.3, was used to measure the strength of the practical relationship between race and outliers/non-outliers.

Kaiser's Measure of Sample Adequacy (MSA) had been computed (Tabachnick & Fidell, 2001) for every exploratory factor analysis. If the MSA was smaller than 0.70 it may have been considered a problem, because the measure can be interpreted with the guidelines given in Section 3.1.2.3.

Skewness and Kurtosis of variables used in the factor analysis were computed to assure that the absolute value does not exceed 2 and 7 respectively, because if a variable is too skew and strongly leptokurtic problems with the factor analysis may occur (see Section 3.1.2.3).

In Chapter 5 a comparison between the constructs retained for the study sample on a specific psychometric test and the sample used when the test was standardised is made.

4.7.3. Methods Used to Address the Third and Fourth Research Questions

As also stated in Chapter 1 the third and fourth research questions in this study are the following:

1. Which of the available reliable and construct valid predictors are the best at predicting academic success for BCom, BSc, BA and BPharm students respectively?
2. Are there models which can adequately predict academic success for each of the BCom, BPharm, BA, and BSc degrees respectively?

Three statistical techniques were used in this study to answer these two questions: CHAID, logistic regression, and predictive discriminant analysis. These three techniques were then used to fit models for the BCom, BSc, BA, and BPharm groups separately. CHAID has been used as an exploratory method to identify important independent variables as well as to identify possible interactions for fitting the models. Logistic regression (which requires fewer assumptions than predictive discriminant analysis) is seen as the main model fitting technique of this study. Discriminant analysis was used to see if the models found in logistic regression were related to those found in discriminant analysis.

The methods used in this study to ensure that **multicollinearity** is not present were to correlate all the continuous predictor variables with one another. If very high absolute values of the correlations occurred (e.g. more than 0.75) between pairs of variables, only one variable of such a pair was chosen (see Section 3.3.2). For instance, matric mathematics performance correlates highly in the BPharm group with *Mscore*. Because *Mscore* are considered more representative of the general academic performance than performance in mathematics only, *Mscore* was entered in the model. The variance inflation factors of all the predictor variables were then computed for each degree type and if a variable's inflation factor was above 5, the variable was not considered for inclusion in a model (see Section 3.2.4). The same precautions which had been taken to rule out **singularity** when doing factor analysis had been taken for model fitting techniques.

4.7.3.1. Criteria for CHAID, Stepwise Logistic Regression and Stepwise Discriminant Analysis

As three techniques were used in this study for model fitting, the objective was to set similar criteria when using these techniques. The reason for doing this was to have a basis for comparing the results yielded by the different model building techniques.

The decision was made to use the stepwise selection as a selection procedure (see Section 3.2.2.8). The criteria for stepwise logistic regression and stepwise predictive discriminant analysis for a predictor to be entered in the model were chosen by keeping the restriction for the number of predictors in the model in mind (see Sections 3.2.2.9 and 3.2.3.2). By experimenting with the criteria values care had been taken that the number of predictors was not too high for the sample sizes. Similar criteria were then set for stepwise logistic regression and stepwise discriminant analysis.

The CHAID analyses in this study were done using SPSS. CHAID was used as an exploratory technique to determine the most significant predictors and to detect possible interactions. The significant level for splitting nodes used in this study was chosen in the same way as for a variable to enter the model in stepwise logistic regression and stepwise discriminant analysis. For merging categories the p-level chosen was the same as for logistic regression for a variable to stay in the model. The scale independent (predictor) variables were all divided into 10 categories. The Pearson chi-square statistic for determining node splitting and merging categories were used. The maximum number of iterations was set as 100. The minimum change in expected cell frequencies was 0.001. The Bonferroni method for adjusted significant values was used in this study (Kass, 1980). As a result of the fact that datasets for model fitting were relatively small the minimum for a parent node was set to be 50 and for a child node 20, unlike the default values used in SPSS, namely 100 for parent node and 50 for child node. The automatic option where the default depth of a tree is 3 was used. No cross validation was done for CHAID, because CHAID was used for exploratory purposes only, as a result of small sample sizes (see Section 3.2.1.4). The decision was also taken that if *Race* were not found to be a significant predictor when using the CHAID procedure, it would then not be entered as an independent variable in the other two model fitting techniques, because of the fact that the race of students was dominantly white and a very small number of students in the study samples were non-white.

4.7.3.2. Logistic Regression

Logistic regression should be seen as the major model fitting technique of this study. SAS was used in this study to perform the stepwise logistic regression analyses. The chi-square values of the Wald tests for the maximum likelihood estimates of the predictors were used to compute the effect sizes. The Hosmer and Lemeshow goodness-of-fit chi-square value was used to calculate the effect size w to determine if the model fitted practically significantly (see Section 3.3.3).

The odds ratios were computed to determine the practical significance of a single predictor in the model. If the odds ratio was less than 1 the reciprocal of the odds ratio was determined for easier interpretation (see Section 3.3.4). To determine the adjusted odds ratio for an interval scaled variable (see Section 3.2.2.7) the unit that was chosen in this study was one standard deviation of the specific predictor variable. One standard deviation is approximately equal to two stanines (Schepers, 1992).

Correlations between the separate selected continuous variables and the logit were computed to determine if a linear relationship between the predictor and the logit existed (see Section 3.2.2.7) to interpret the odds ratio correctly. A best subset selection was also made (see Section 3.2.2.8). Cross validation was done by using the leave-one-out method and a classification table was constructed to determine the sensitivity, specificity and observed hit rate. The area under the ROC curve was computed to determine the discrimination ability of the model.

4.7.3.3. Predictive Discriminant Analysis

The joint distribution of the variables used for the stepwise selection procedure in discriminant analysis has to be multinormal according to the assumption in Section 3.2.3.1. The conclusion of multinormality could only be made if each separate predictor was normally distributed, which was not the case in this study. However, according to McLachlan (2004) linear discriminant analysis is fairly robust against departure from assumptions.

Similar criteria for stepwise predictive discriminant analysis as for logistic regression were used in this study. SAS was also used. The priors used were the expected hit rate for every different degree type. In this study a chi-square test was used to determine if the covariance matrices were equal. The effect size w was then calculated (see Section 3.3.3) to decide if a linear rule or quadratic rule could be used. The linear or quadratic discriminant functions based on either rule were then obtained.

Cross validation was done by using the leave-one-out method and a classification table was constructed to determine the sensitivity, specificity and observed hit rate.

4.7.3.4. Improvement over Chance

The effect size indices, I, for improvement over chance for both the stepwise logistic regression and the stepwise discriminant analysis were computed to determine if the models found were better than chance (see Section 3.3.5).

Chapter 5

5. Reliability and Validity of Psychometric Tests

This chapter addresses the first two research questions: Are the SAT 78, GSAT, SSHA, PHSF, and 19 FII;

1. reliable instruments for the study sample and, if so, how does the reliability of these instruments compare with the reliability at the time of standardisation?
2. construct valid instruments for the study sample and, if so, how does the construct validity of these instruments compare with the construct validity at the time of their standardisation?

As stated in Chapter 4 Cronbach alpha coefficients had been computed to determine the reliability of the subtests of the psychometric tests and exploratory factor analyses were done to examine the construct validity. The results are discussed in the next sections and evaluations are made.

5.1. Introduction

The reliability of a test was defined and discussed in Chapter 2. The reliability of a test refers to the consistency of scores obtained by the same persons when they are re-examined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions (Anastasi & Urbina, 1997).

The validity of a test concerns what the test measures and how well it does so (Anastasi & Urbina, 1997). If valid it measures then what it is supposed to measure.

Because “reliability is a characteristic of data” (Eason, 1991: 84) researchers must attend to the influence that the participants themselves have on score quality in every study. As Thompson (1994: 839) explained, because total score variance is an important aspect of reliability, the participants involved in the study will themselves affect score reliability: “the same measure, when administered to more heterogeneous or more homogenous sets of subjects, will yield scores with differing reliability”.

Given the diversity of participants across studies, constructors of every study involving psychometric tests should provide reliability coefficients on the scores for the data analysed. As Pedhazur and Schmelkin (1991: 86) have argued: “Researchers who bother at all to

report reliability estimates for the instruments they use (many do not) frequently report only reliability estimates contained in the manuals of the instruments or estimates reported by other researchers. Such information may be useful for comparative purposes, but it is imperative to recognise that the relevant reliability estimate is the one obtained for the sample used in the study under consideration".

The same argument holds for validity. Validity was defined and discussed in Chapter 2. According to Nunnally and Bernstein (1994) validity is a matter of degree rather than an all or none property and validation is an unending process. Most psychological measures need to be constantly evaluated and re-evaluated to see if they are behaving as they should. The most precise and efficient measures are those with established construct validity (Clark & Watson, 1995).

In the light of this it is meaningful to ask the question of how reliable and construct valid the SAT 78, GSAT, PHSF, SSHA, and 19 FII are on a group of students some years after their initial standardisation.

5.2. Study Sample

None of the testees took all five tests. However, all the testees were students of, or applicants to, the same university where students are predominantly white Afrikaans speaking and from formerly advantaged communities.

The majority of testees were around 18 years of age and either undertook the test in the September preceding entry to university, or in the January of their first year, before commencing their studies.

According to Nunnally and Bernstein (1994) it is possible to combine the samples from the different year groups of each test into one study sample because the respondents are from the same background, culture, and age and the test was conducted at the same time of the year for each year group. The psychometric tests' data for the year groups 2003 to 2007 were combined to form the Study Sample (2003-2007) for which the reliability and construct validity were estimated.

Applications to Pharmacy were required to do the GSAT, SSHA, PHSF, and 19 FII. Undertaking psychometric tests was not compulsory for any other students but if they chose to do them they were required to complete the SAT, SSHA, PHSF, and 19 FII. No random

selection of students was made at any time. From of an academic point of view the respondents were a selected group as a result of the fact that they had either passed grade 11 (most of the Pharmacy applicants) or were already selected to enter the university.

5.3. SAT 78

5.3.1. Study Sample

The total number of respondents was 2 084. Thus the requirement for **sample size** is more than adequately met (see Section 3.1.2.3).

The descriptive statistics for the subtests of the SAT 78 on the Study Sample (2003-2007) are given in Table 5.1.

Table 5.1 Descriptive statistics of the SAT 78 subtests on the Study Sample (2003-2007)

Subtest	*Mean*	*Std dev*	*Skewness*	*Kurtosis*	*Min*	*Max*	*Domain*
Verbal Comprehension	17.37	4.11	-0.28	3.05	0	28	0 - 30
Calculations	16.23	6.33	0.50	3.25	0	40	0 - 40
Disguised Words	16.25	5.71	0.03	2.43	0	30	0 - 30
Comparison	19.66	3.85	-0.36	3.73	0	30	0 - 30
Pattern Completion	17.70	5.59	0.08	2.30	3	30	0 - 30
Figure Series	4.05	8.14	1.68	4.19	0	30	0 - 30
Spatial 2D	19.00	7.11	-0.58	2.63	0	30	0 - 30
Spatial 3D	19.23	5.60	-0.65	3.23	1	30	0 - 30
Memory (Paragraph)	11.92	4.05	-0.16	2.35	1	20	0 - 20
Memory (Symbols)	24.74	4.87	-1.01	3.59	4	30	0 - 30

n = 2 084

5.3.2. Reliability

When the SAT 78 was standardised reliability coefficients of the subtests of the SAT 78 were calculated using the Kuder-Richardson formula 8 (K-R 8) for subtests 1 to 10. In this study Cronbach alpha coefficients were calculated in SAS to determine internal reliability. The

Kuder-Richardson formula 8 is a special case for the Cronbach alpha coefficient (Anastasi & Urbina, 1997).

The original reliability coefficients of the subtests for the sample on which the test was standardised in 1978 as well as the reliability coefficients for our sample are given in Table 5.2. As stated in Chapter 4 reliability coefficients of 0.70 and above are necessary.

Table 5.2 Reliability Coefficients of SAT 78 Subtests on the 1978 Sample and Study Sample (2003-2007)

Subtest	*K-R 8 (1978)*[1]	*Cronbach Alpha Coefficient (2003-2007)*
Verbal Comprehension	0.717	0.725
Calculations	0.921	0.903
Disguised Words	0.788	0.833
Comparison	0.762	0.773
Pattern Completion	0.834	0.849
Figure Series	0.852	0.983
Spatial 2D	0.918	0.928
Spatial 3D	0.838	0.850
Memory (Paragraph)	0.762	0.771
Memory (Symbols)	0.836	0.853

[1] From Fouche & Verwey, 1978, n = 1 453

Satisfactory to high Cronbach alpha coefficients were obtained in this study and the values compare very well with the K-R 8 values obtained in 1978 for the different subtests. This means that the SAT 78 is still a reliable test for the study sample.

5.3.3. Construct Validity

The SAT 78 was standardised using the following procedures: Construct validity was calculated through exploratory factor analysis and predictive validity was measured by calculating correlation coefficients between SAT 78 subtests and certain examination marks the testees obtained in different school subjects at a certain time (Fouche & Verwey, 1978).

The chi-square value at significance level 0.001, with 10 degrees of freedom (that is the number of subtests of the SAT 78) is 29.59. The threshold value h_{ii} according to Equation 3.16 is then 0.02 (see Section 3.1.2.3). Thus, if a respondent had a h_{ii} value higher than 0.02 it was classified as an outlier. None of the h_{ii} values were higher than 0.02 and hence none of the cases (respondents) was identified as **outliers.**

In this study the main focus was on construct validity and no predictive validity was done at this stage of the study. A principal component exploratory factor analysis using SAS was done to determine construct validity. The scores of the 10 subtests mentioned were the variables used in the factor analysis. A varimax rotation was done. The overall measure of sampling adequacy (**MSA**) was 0.85 which, as stated in Section 3.1.2.3, is meritorious. This means that the correlation matrix was appropriate for a factor analysis (Hair *et al.*, 1998).

The maximum absolute value of all the subtests' **skewness** was 1.68 and thus less than 2, while the absolute value of all the subtests' **kurtosis** was smaller than 7 and thus the requirements in this regard for the factor analysis were met.

The factor loadings of the subtests forming the four constructs of the SAT 78 are broken down by language and gender are in Table 5.3.

Table 5.3 Original constructs and factor loadings of the subtests of the SAT 78

Constructs	*Factor Loadings*			
	Afrikaans		English	
	Boys	Girls	Boys	Girls
Construct 1 (Verbal Ability)				
Verbal Comprehension	0.62	0.66	0.52	0.52
Disguised Words	0.64	0.66	0.44	0.47
Memory (Paragraph)	0.43	0.39	0.45	0.38
Construct 2 (Numerical Ability)				
Calculations	0.55	0.64	0.66	0.62
Comparison	0.19	0.50	0.25	0.59

Construct 3 (Visual-Spatial Reasoning)				
Pattern Completion	0.51	0.54	0.49	0.52
Figure Series	0.58	0.61	0.44	0.49
Spatial 2D	0.62	0.60	0.68	0.49
Spatial 3D	0.72	0.60	0.71	0.67
Construct 4 (Memory)				
Memory (Paragraph)	0.46	0.47	0.36	0.45
Memory (Symbols)	0.37	0.60	0.45	0.52

From Fouche & Verwey, 1978, n = 1 453

The factor loadings calculated for the Study Sample (2003-2007) are given in Table 5.4.

Table 5.4 Constructs and factor loadings of the SAT 78 for the Study Sample (2003-2007)

Subtest	*Factor Loadings*		
	Construct 1	Construct 2	Construct 3
Spatial 3D	0.83		
Spatial 2D	0.82		
Pattern Completion	0.66		
Calculations	0.53	0.34	
Memory (Paragraph)		0.80	
Disguised Words		0.69	
Verbal Comprehension	0.52	0.61	
Memory (Symbols)		0.61	
Comparison		0.56	
Figure Series			0.92

Values of 0.3 and above are reported

As can be seen in Table 5.4, the 1978 exploratory factor analysis yielded four significant factors. The SAT 78 manual does not state what portion of variation in the data was explained by these constructs.

In contrast only three significant factors (which again form the constructs) were found in this study. The constructs included different subtests with different factor loadings and were hence largely incomparable to the original constructs. These three constructs explained 59% of the variation in the data. Communalities varied between from 0.43 to 0.87.

These results for the exploratory factor analysis, which were very different to those of the Fouche and Verwey (1978) study, were obtained even though all the conditions, i.e. the MSA, sample size, and skewness, were more than satisfactorily met.

5.3.4. Evaluation

All the subtests of the SAT 78 on the study sample were reliable. One implication is that a reliable estimated IQ can be calculated for each respondent of the study sample. The scores of the 10 subtests may also be used to try to predict academic success by using statistical models such as logistic regression or discriminant analysis. Their predictive validity may also be investigated.

Because the constructs formed from the study sample are not comparable to the original constructs of the SAT 78 it would not make sense to interpret the original constructs for the study sample. For example, a high score in the SAT 78 construct *Numerical Ability* has no meaning for a student in the study sample and should not be considered when deciding which area of study a student should enter or to predict the student's success in a certain field.

5.4. GSAT

5.4.1. Study Sample

The total number of respondents was 591. Thus the requirement for **sample size** is met (see Section 3.1.2.3).

From an academic point of view the testees were an even more select group because they must have had Mathematics and Science as subjects in matric to qualify for admission for the BPharm degree.

Table 5.5 shows descriptive statistics of the GSAT subtests for the study sample.

Table 5.5 Descriptive statistics of the GSAT constructs and subtests on the Study Sample (2003-2007).

Construct	*Mean*	*Standard Deviation*	*Skewness*	*Kurtosis*	*Min*	*Max*	*Domain*
Verbal	**39.10**	**5.31**	**-1.32**	**6.37**	**11**	**49**	**0 - 50**
Word Analogies	19.54	2.89	-1.31	6.55	4	25	0 - 25
Verbal Reasoning	19.56	3.05	-0.61	3.43	0	25	0 - 25
Nonverbal	**37.39**	**5.15**	**-0.72**	**4.09**	**13**	**49**	**0 - 50**
Figure series	18.56	3.05	-1.09	5.96	8	25	0 - 25
Pattern Completion	18.82	3.03	-0.78	4.19	3	25	0 - 25
Total	**76.49**	**9.32**	**-1.17**	**5.88**	**24**	**97**	**0 - 100**

n = 591

5.4.2. Reliability

When the GSAT was standardised reliability coefficients for the *Verbal, Nonverbal*, and *Total* constructs were calculated using the Kuder-Richardson formula 8 (K-R 8). In this study calculated Cronbach alpha coefficients were calculated to determine the constructs' but also the subsets' internal reliability. The reason for calculating reliability for the constructs, was because it was done at the time of standardisation, and the objective of this study is to come as close as possible with procedures as reported in the manuals of the different tests.

The reliability coefficients of constructs for the sample in 1991 and Study Sample (2003-2007) as well the coefficients for the subtests for the Study Sample (2003-2007) are given in Table 5.6.

High Cronbach alpha coefficients were obtained in this study for both the subtests and the constructs. The constructs' reliability coefficients compare very well with the K-R 8's obtained in 1991.

As all the Cronbach alpha coefficients were above 0.70 and the values obtained compare favourably with the K-R 8 coefficients obtained in the validation study in 1991. The GSAT remains a reliable tests.

Table 5.6 Reliability coefficients of GSAT constructs on the 1991 Sample and for constructs and subtests on the Study Sample 2003-2007

Construct	K-R 8 (1991)[1]	Cronbach Alpha Coefficient (2003-2007)
Verbal	**0.91**	**0.81**
Word Analogies		0.71
Verbal Reasoning		0.69
Nonverbal	**0.91**	**0.77**
Figure Series		0.68
Pattern Completion		0.69
Total	**0.95**	**0.87**

[1] From Claassen *et al.*, 1991. Subtest reliability coefficients were not reported, n = 138 (grade 12)

5.4.3. Construct Validity

When the GSAT was standardised construct validity was determined by exploratory factor analysis and predictive validity was determined by calculating correlation coefficients between GSAT subtests and certain examination marks the testees obtained in different school subjects at a certain time (Claassen *et al.*, 1991). In this study the main focus was on construct validity and predictive validity was not investigated. A principal component exploratory factor analysis was done to determine construct validity. The scores of the 4 subtests were the variables used in the factor analyses.

The chi-square value at significance level 0.001, with 4 degrees of freedom (that is the number of subtests of the GSAT) is 18.47. The threshold value h_{ii} according to Equation 3.16 is then 0.03 (see Section 3.1.2.3). Thus, if a respondent had a h_{ii} value higher than 0.03 it was classified as an outlier. With this criterion, 7 out of the 591 respondents were identified as **outliers**. A contingency table with race versus outlier was created to determine if there was a relationship between race and being an outlier. The relationship between race and outliers was of a medium effect (see Section 3.3.3). Still, only 4 of the 12 black students were outliers as can be seen in the following table. This is such a small number that the effect on reliability and construct validity is likely to be negligible. In Table 5.7 the numbers and percentages in each cell are given.

Table 5.7 GSAT testees and outliers by race

Race	*Testees*		*Outliers*		
	Number	% of Total	Number	% of Testees	% Outliers
Asian	11	2%	0	0%	0%
Black	12	2%	4	33%	57%
Coloured	10	2%	1	10%	14%
White	528	89%	2	0%	29%
Unknown	30	5%	0	0%	0%
Total	**591**	**100%**	**7**	**1%**	**100%**

The overall measure of sampling adequacy (**MSA**) was 0.75 which, as stated in Section 3.1.2.3, is *middling*. This is high enough to indicate that a factor analysis was appropriate.

The maximum absolute value of **skewness** of a subtest was 1.3 which is lower than maximum acceptable value of 2 and the absolute value of the **kurtosis** of all the subtests was also below 7 (see Section 3.1.2.3).

Testees at this university completed the abbreviated speed loaded version of the GSAT. The GSAT manual does not report the factor loadings of the subtests on the constructs for this version. Instead, it lists the factor loading for the full version GSAT.

The factor loadings of the full version GSAT on the only significant principal component factor, which forms the only construct, are given in Table 5.8 for the original sample in 1991.

Table 5.8 Factor loadings of the original subtests of the full version GSAT 78 on the first principal component (1991)

Subtests	Factor Loading
Word Analogies	0.83
Word Pairs	0.85
Verbal Reasoning	0.89
Figure Series	0.83
Pattern Completion	0.82
Figure Analogies	0.85

From Claassen *et al.,* 1991, n = 786

The factor explained 50% of the variation in the data. The results of this factor analysis imply a strong common factor which is most probably Spearman's ***g*** (Claassen *et al.* 1991; Nunnally & Bernstein, 1994). The factor loadings calculated for the Study Sample (2003-2007) are given in Table 5.9.

Table 5.9 Factor loadings on first principal component for the Study Sample (2003-2007)

Subtest	Factor Loading
Verbal Reasoning	0.86
Word Analogies	0.76
Figure Series	0.78
Pattern Completion	0.71

n = 591

The exploratory factor analysis yielded only one significant construct for Study Sample (2003-2007) which is similar to the factor pattern for the original sample in 1991.

According to Stevens (1992) a factor loading of 0.7 is considered very high. The construct found for the study sample is again most probably Spearman's ***g***. The construct explained 60% of the variation in the data. Communalities varied from 0.50 to 0.73.

The constructors of the GSAT investigated the possibility of separating verbal and nonverbal intelligence into two constructs. To do this they specified a two factor structure and performed a quartimin rotation on the data. Table 5.10 shows the factor loadings generated using this method.

In this factor analysis all the subtests loaded very highly on the first factor and none of the subtests loaded with more than the required 0.3 on the second factor. The second factor explained very little of the variation in the data and the correlation between the two factors was 0.03, which is very low. This means that it was not possible to distinguish between a verbal and nonverbal factor (Claassen *et al.*, 1991).

It was not possible to specify a two factor oblique rotation because it is not feasible to obtain two valid factors from only four variables (Schepers, 1990).

Table 5.10 Two factor structure factor loadings (1991)

Subtest	*Construct 1*	*Construct 2*
Word Analogies	0.78	0.26
Word Pairs	0.83	0.25
Verbal Reasoning	0.85	0.06
Figure Series	0.81	-0.05
Pattern Completion	0.84	-0.23
Figure Analogies	0.83	-0.24

n = 939

5.4.4. Evaluation

This study revealed that all four subtests, as well as the *Verbal*, *Nonverbal*, and *Total* aptitude scores, of this version of the GSAT for the study sample were reliable. The scores of these three aptitudes and the four subtests may be used to try to predict academic success by using logistic regression or discriminant analysis. Another implication is that these tests may be used to determine their predictive validity when correlated with certain academic marks obtained by the testees during their academic courses. It can also be used in the selection process for the BPharm degree at this university because high scores would be an indication of academic intelligence (Claassen *et al.*, 1991).

This study also indicates that this version of the GSAT is construct valid for the study sample. That means that it measures the academic intelligence and scholastic aptitude of the prospective pharmacy students accurately. As a result of the fact that only one construct was extracted when doing a principal components factor analysis in this study it was also not possible to distinguish between a verbal and nonverbal construct for the study sample. This construct distinction was also not present at the time of standardisation on the sample in 1991 although it is a traditional distinction in intelligence tests (Claassen *et al.*, 1991).

5.5. SSHA

5.5.1. Study Sample

The total number of respondents was 2 590. Thus, the requirement for **sample size** is more than adequately met (see Section 3.1.2.3).

The descriptive statistics for the subtests of the SSHA on the Study Sample (2003-2007) are given in Table 5.11.

Table 5.11 Descriptive statistics of SSHA subtests for the Study Sample (2003-2007)

Subtest	Mean	Standard Deviation	Skewness	Kurtosis	Min	Max	Domain
Delay Avoidance	21.57	8.94	0.22	6.37	1	48	0 – 50
Work Methods	24.29	8.47	0.04	6.55	2	50	0 – 50
Study Habits	45.86	16.13	0.16	3.43	5	97	0 – 100
Teacher Approval	24.72	8.66	-0.03	4.09	1	50	0 – 50
Education Acceptance	25.87	7.81	-0.12	5.96	0	47	0 – 50
Study Attitude	50.59	15.47	-0.10	4.19	5	94	0 – 100
Study Orientation	96.45	29.50	0.03	5.88	17	184	0 – 200

n = 2 590

5.5.2. Reliability

When the SSHA was standardised reliability coefficients of the subtests of the SSHA were calculated using split half reliability coefficients. In this study Cronbach alpha coefficients were calculated to determine internal reliability.

The reliability coefficients of subtests for the samples in 1974 and in this study are given in Table 5.12. The manual does not report split-half coefficients for *Study Habits*, *Study Attitude*, and *Study Orientation*.

High Cronbach alpha coefficients were obtained in this study and the values compare very well with the split-half coefficients obtained in 1974 for the different subtests of the SSHA. This means that the SSHA is a highly reliable test for this study sample.

Table 5.12 Reliability coefficients of the SSHA subtests on the Study Sample (2003-2007)

Subtest	Spit-half Coefficient (1974) [1]	Cronbach Alpha Coefficient (2003-2007)
Delay Avoidance	0.833	0.85
Work Methods	0.835	0.84
Study Habits		0.91
Teacher Approval	0.873	0.87
Education Acceptance	0.805	0.81
Study Attitude		0.91
Study Orientation		0.95

[1] From du Toit, 1974, n = 2 790

5.5.3. Construct Validity

Unlike with the SAT 78 and GSAT it seems from the SSHA's manual that construct validity was not calculated through exploratory factor analysis. The SSHA's predictive validity was determined by calculating correlation coefficients between the subtests and the average of the testees' examination marks (du Toit, 1974).

The chi-square value at significance level 0.001, with 4 degrees of freedom (that is the number of subtests of the SSHA) is 18.47. The threshold value h_{ii} according to Equation 3.16 is then 0.01 (see Section 3.1.2.3). Thus, if a respondent had a h_{ii} value higher than 0.01 it was classified as an outlier. Only one respondent has a h_{ii} value of higher than 0.01 and hence only one of the cases (respondents) was identified as an **outlier.**

In this study the main focus was on construct validity and no predictive validity was done. A principal component exploratory factor analysis using SAS was done to determine construct validity. The scores of the 4 subtests *Delay Avoidance*, *Work Methods*, *Teacher Approval*, and *Education Acceptance* were used in the factor analysis. The overall measure of sampling adequacy (**MSA**) was 0.79 which, as stated in Section 3.1.2.3, is *middling*. This means that the correlation matrix was appropriate for a factor analysis (Hair *et al.*, 1998)

The maximum absolute value of **skewness** was 0.22 which is less than 2 and the absolute values for the **kurtosis** of the subtests were less than 7 and hence not a problem for the factor analysis (see Section 3.1.2.3). The factor loadings calculated for the Study Sample (2003-2007) are given in Table 5.13.

Table 5.13 Factor loadings on first principal component on the Study Sample (2003-2007)

Subtest	Factor Loadings
Delay Avoidance	0.86
Work Methods	0.87
Teacher Approval	0.84
Education Acceptance	0.92

In this factor analysis all the subtests loaded very highly on the first factor. The factor explained 76% of the variation in the data. Communalities varied from 0.70 to 0.85.

5.5.4. Evaluation

This study revealed that all the subtests of the SSHA scores for the Study Sample (2003-2007) are reliable subtests. The scores of these subtests may be used to try to predict academic success by using logistic regression or discriminant analysis. Their predictive validity can also be investigated in the future. It can also be used in the selection process for the BPharm degree at this university because high scores would be an indication of a student with good academic performance (du Toit, 1974). Obtaining high Cronbach alpha values on this questionnaire for the different subtests on this study sample contradicts the findings of Penny (1984). Low reliability coefficients with two different methods, namely split-half and an item analysis were obtained on all four subscales of the SSHA by Penny (1984) for the study sample used at that time.

This study also indicates that this version of the SSHA is construct valid for this study sample. In other words, it measures the study orientation of the participants. As only one construct was extracted when doing a principal components factor analysis in this study it was not possible to distinguish between all the different facets of study habits.

5.6. PHSF

5.6.1. Study Sample

The study sample for the PHSF was the same as for the SSHA as explained in Section 5.5.1 with n = 2 585. Thus the requirement for **sample size** is more than adequately met (see Section 3.1.2.3). The descriptive statistics for the subtests of the PHSF on the study sample are given in Table 5.14.

Table 5.14 Descriptive statistics of PHSF subtests for the Study Sample (2003-2007)

Subtest	Mean	Standard Deviation	Skewness	Kurtosis	Min	Max	Domain
Self-confidence	29.19	5.87	-0.14	3.11	8	45	0 – 45
Family Influences	31.40	7.05	-0.17	3.40	4	45	0 – 45
Moral Sense	33.35	5.79	-0.31	2.67	11	45	0 – 45
Health	31.68	6.04	-0.40	3.09	7	45	0 – 45
Self-esteem	25.95	5.80	-0.30	3.21	3	43	0 – 45
Sociability-S	28.98	7.32	-0.41	3.27	0	45	0 – 45
Formal Relations	30.55	5.29	0.01	2.96	12	45	0 – 45
Nervousness	25.66	5.67	0.17	2.98	3	45	0 – 45
Personal Freedom	34.87	6.86	-0.95	3.97	3	45	0 – 45
Self-control	27.29	5.00	-0.11	3.28	4	44	0 – 45
Sociability-G	28.33	7.43	-0.45	3.18	2	45	0 – 45
Desirability Scale	17.03	4.92	0.10	3.14	3	36	0 – 45

n = 2 585

5.6.2. Reliability

Reliability coefficients for PHSF subtests were calculated in 1983 using the split-half method. As said in this study Cronbach alpha coefficients were calculated to determine internal reliability. It can be shown mathematically that Cronbach alpha coefficients or Kuder-Richardson coefficients (which are similar) are the mean of all split-half coefficients resulting from different splitting (Anastasi & Urbina, 1997).

The reliability coefficients of subtests for the samples in 1983 and in this study are given in Table 5.15. In this study no reporting of reliability coefficients by gender is done.

Table 5.15 Reliability coefficients of the PHSF subtests

Subtest	Spit-half Coefficient (1983)		Cronbach Alpha Coefficient (2003-2007)
	Boys	Girls	
Self-confidence	0.80	0.79	0.83
Self-esteem	0.75	0.74	0.79
Self-control	0.71	0.70	0.66
Nervousness	0.74	0.74	0.70
Health	0.80	0.85	0.82
Family Influences	0.85	0.88	0.87
Personal Freedom	0.87	0.89	0.87
Sociability-G	0.88	0.89	0.89
Sociability-S	0.91	0.89	0.87
Moral Sense	0.79	0.77	0.80
Formal Relations	0.83	0.80	0.80
Desirability Scale	0.75	0.78	0.68

From Fouche & Grobbelaar, 1983, n (boys) = 909, n (girls) = 879

Satisfactory to high Cronbach alpha coefficients were obtained in this study and the values compare very well with the split-half values obtained in 1983 for the different subtests. This means that the PHSF remains a reliable test for this study sample.

5.6.3. Construct Validity

The PHSF's construct validity was determined through exploratory factor analysis in 1983. In this study a principal component exploratory factor analysis using SAS was again done to determine construct validity. The scores of the 12 subtests in Table 5.15 were the variables used in the factor analysis. A varimax rotation was done.

The chi-square value at significance level 0.001, with 12 degrees of freedom (that is the number of subtests of the PHFS) is 32.91. The threshold value h_{ii} according to Equation 3.16 is then 0.01 (see Section 3.1.2.3). With this criterion 133 respondents were identified as **outliers**. A contingency table with race versus outlier was created and the phi coefficient was determined as 0.09 which meant that there was no practically significant relationship between race and outliers in the data. The large number of outliers may be explained by the diverse nature of the study sample.

The overall measure of sampling adequacy (**MSA**) was 0.83 which, as stated in Section 3.1.2.3, is *meritorious*. This means that the correlation matrix was appropriate for a factor analysis (Hair *et al.*, 1998).

The maximum absolute value of **skewness** was 0.95 which is less than 2 and the absolute values of the **kurtosis** of the subtests were also below 7 and hence not a problem for the factor analysis (see Section 3.1.2.3).

The factor groupings with highest loadings from PHSF manual are reported in Table 5.16. The manual neither reports the factor loadings nor the number of the sample size.

Table 5.16 Constructs and subtests of PHSF (1983)

Construct 1 (Nervousness)
Nervousness
Health
Self-esteem
Self-control
Construct 2 (Home Relations)
Family Influences
Personal Freedom
Construct 3 (Moral Sense)
Moral Sense
Desirability Scale
Formal Relations
Self-control
Construct 4 (Sociability)
Sociability-G
Sociability-S
Construct 5 (Self-confidence)

Self-confidence	
Formal Relations	
Self-esteem	
Construct 6 (School Relations)	
Formal Relations	
Construct 7 (Self-esteem)	
Self-esteem	
Construct 8 (Personal Freedom)	
Personal Freedom	

From Fouche & Grobbelaar, 1983

The factor loadings for the Study Sample (2003-2007) are given in Table 5.17.

Table 5.17 Rotated factor pattern for Study Sample (2003-2007)

Subtest	Construct 1*	Construct 2*	Construct 3*
Self-control	0.77		
Moral Sense	0.74		
Nervousness	0.59	0.31	
Formal Relations	0.59	0.43	
Desirability	-0.72		
Sociability-G		0.80	
Self-confidence	0.38	0.74	
Self-esteem		0.72	
Sociability-S	-0.39	0.65	
Health	0.43	0.44	
Personal Freedom			0.86
Family Influences			0.82

*Absolute loadings < 0.3 not reported

A very different factor pattern was detected for this study sample than the factor pattern retained at the time when the PHSF was standardised. Eight constructs (factors) were retained in 1983. Three constructs (factors) were retained when working with this study sample. These three factors explained 61% of the variation in the data. Communalities varied from 0.47 to 0.78.

These results for the exploratory factor analysis, which were very different to those of the Fouche and Grobbelaar (1983) study, were obtained even though all the conditions concerning the sample size, MSA, skewness, and kurtosis were more than satisfactorily met.

5.6.4. Evaluation

This study revealed that all the subtests of the PHSF on the study sample were reliable. *Desirability* was not eligible for use in the prediction models done in this study because it is only used by the PHSF for validation purposes. The remaining 11 subtests can be used to determine predictive validity when correlated with certain academic marks obtained by the respondents during their future academic courses. The scores of these 11 subtests may also be used to try to predict academic success by using logistic regression or a discriminant analysis.

However, keeping the meaning of construct validity in mind, which is namely that a test must measure what it is supposed to measure, serious doubts about the construct validity of the PHSF for this study sample are raised. This means that it would not be wise to interpret an individual's score for a specific type of adjustment and guide students on behalf of it to a future career. It would also not be fair to select an applicant for a specific field of study solely on the grounds of the results of this questionnaire.

5.7. 19 FII

5.7.1. Study Sample

The study sample for the 19 FII was the same as for the SSHA as explained in Section 5.5.1 with n = 2 597. The number of testees, however, differs by a small number for each test because of data gathering and cleansing issues.

The descriptive statistics for the subtests of the 19 FII on the study sample are given in Table 5.18.

Table 5.18 Descriptive statistics of 19 FII subtests for the Study Sample (2003-2007)

Subtest	Mean	Standard Deviation	Skewness	Kurtosis	Min	Max	Domain
Fine Arts	17.94	11.42	0.41	2.31	0	45	0 – 45
Performing Arts	12.43	11.87	0.84	2.76	0	45	0 – 45
Language	13.83	11.57	0.73	2.64	0	45	0 – 45
Historical	15.70	10.94	0.60	2.56	0	45	0 – 45
Service	13.42	8.28	0.54	2.90	0	45	0 – 45
Social Work	19.53	11.65	0.26	2.27	0	45	0 – 45
Sociability	33.26	9.32	-0.82	3.34	0	45	0 – 45
Public Speaking	15.87	11.48	0.55	2.52	0	45	0 – 45
Law	16.15	13.25	0.52	2.17	0	45	0 – 45
Creative Thought	31.12	8.92	-0.53	3.03	0	45	0 – 45
Science	18.71	12.89	0.24	1.87	0	45	0 – 45
Practical-Male	15.81	12.96	0.58	2.34	0	45	0 – 45
Practical-Female	15.62	9.72	0.40	2.50	0	45	0 – 45
Numerical	21.12	12.40	-0.07	1.98	0	45	0 – 45
Business	23.80	12.43	-0.07	2.03	0	45	0 – 45
Clerical	15.36	10.44	0.57	2.68	0	45	0 – 45
Travel	26.72	10.35	-0.28	2.37	0	45	0 – 45
Nature	11.85	11.87	0.88	2.74	0	45	0 – 45
Sport	19.30	11.82	0.22	2.17	0	45	0 – 45
Work-Hobby	15.19	2.58	-1.18	7.27	0	20	0 – 20
Active-Passive	10.59	3.16	-0.13	3.15	0	20	0 – 20

n = 2 597

5.7.2. Reliability

Reliability coefficients for 19 FII subtests were calculated in 1977 using the split-half method. Cronbach alpha coefficients were calculated to determine internal reliability on the study sample for this test.

The reliability coefficients of subtests for the samples in 1977 and in this study are given in Table 5.19. In this study reliability coefficients by gender are not reported.

Table 5.19 Reliability coefficients of the 19 FII subtests

Subtest	Spit-half Coefficient (1983)		Cronbach Alpha Coefficient (2003-2007)
	Boys	Girls	
Fine Arts	0.97	0.97	0.95
Performing Arts	0.95	0.97	0.95
Language	0.94	0.95	0.96
Historical	0.94	0.94	0.94
Service	0.92	0.9	0.89
Social Work	0.96	0.96	0.96
Sociability	0.96	0.95	0.95
Public Speaking	0.97	0.97	0.95
Law	0.98	0.98	0.98
Creative Thought	0.96	0.95	0.94
Science	0.97	0.95	0.96
Practical-Male	0.98	0.97	0.98
Practical-Female	0.96	0.96	0.98
Numerical	0.97	0.97	0.96
Business	0.98	0.97	0.96
Clerical	0.96	0.97	0.95
Travel	0.92	0.93	0.94
Nature	0.97	0.97	0.97
Sport	0.95	0.96	0.94
Work-Hobby	0.81	0.75	0.54
Active-Passive	0.73	0.68	0.59

From Fouche & Alberts 1977, n (boys) = 408, n (girls) = 495

High Cronbach alpha coefficients were obtained in this study for the 19 subtests of the 19 FII which compare very well with the split-half values obtained in 1977 for the different subtests.

For the two additional tests, *Work-Hobby* and *Active Passive,* the Cronbach alpha values were below 0.7 which means that these two additional subtests are not reliable on the Study Sample (2003-2007). This means that the 19 main subtests of the 19 FII remain as reliable tests on this study sample.

5.7.3. Construct Validity

No method was mentioned in the manual how construct validity was determined for the 19 FII in 1977.

The chi-square value at significance level 0.001, with 19 degrees of freedom (that is the number of subtests of the 19 FII) is 43.82. The threshold value h_{ii} according to Equation 3.16 is then 0.02 (see Section 3.1.2.3). With this criterion 8 respondents were identified as **outliers**. A contingency table with race versus outlier was created and the phi coefficient was determined as 0.04 which means that there is no practically significant relationship between race and outliers in the data.

In this study the main focus was on construct validity. A principal component exploratory factor analysis was done to determine construct validity. The scores of the 19 subtests mentioned were the variables used in the factor analysis. A varimax rotation was done. The overall measure of sampling adequacy (**MSA**) was 0.75 (Hair *et al.*, 1998). This means that the correlation matrix was appropriate for a factor analysis.

The maximum absolute value of **skewness** was 1.18 which is less than 2 and the absolute value of the **kurtosis** of the subtests were less than 7 and hence not a problem for the factor analysis (see Section 3.1.2.3). The kurtosis of *Work-hobby* is 7.27, but this subtest was not used as a variable in the factor analysis, because it was not used as a variable at the time of standardisation. However, the reliability of this variable for the Study Sample (2003-2007) was lower than 0.7 which means that it could not be used as a reliable variable in the factor analysis.

The factor groupings 19 FII manual are reported in Table 5.20.

Table 5.20 Interest factors mentioned in 19 FII manual

Interest Factors	*Boys*	*Girls*
Fine Arts	Fine Arts	Fine Arts Creative Thought
Performing Arts	Performing Arts	Performing Arts
Language	Language	Language Historical
Service	Service Clerical Public Speaking	Service Clerical Public Speaking
Social Work	Social Work	Social Work
Sociability	Sociability Business Travel Sport	Sociability Business Travel Sport Service
Manipulation of scientific principles	Science	Science
Influencing the ideas and thinking of others	Public Speaking Law Language	
Manipulation of own thoughts and ideas	Creative Thought Science Nature	
Manipulation of thoughts and ideas		Public Speaking Law Creative Thought Business Language
Manipulation of things	Practical Male	Practical Male Nature
Manipulation of figures	Nature	Nature
Nature	Nature	
Evasion of occupational responsibility	Active-Passive Work-Hobby	
Travel	Travel Historical	

From Fouche & Alberts, 1977

The factor loadings for the Study Sample (2003-2007) are given in Table 5.21.

Table 5.21 Subtests and their factor loadings for Study Sample (2003-2007)

Subtest	*Factor Loadings*					
	Construct 1	**Construct 2**	**Construct 3**	**Construct 4**	**Construct 5**	**Construct 6**
Language	0.74		0.42			
Fine Arts	0.72					
Performing Arts	0.69					
Practical-Female	0.57			0.30		0.40
Historical	0.50	0.45	0.37			
Social Work	0.43					0.33
Nature		0.84				
Practical-Male		0.75				
Public Speaking			0.80			
Law			0.78			
Business	-0.33	0.42	0.49			
Sociability				0.83		
Travel				0.61		
Sport		0.48		0.53		
Creative Thought					0.78	
Science					0.72	
Numerical	-.0.34				0.68	0.36
Clerical	0.40					0.88
Service				0.39		0.63

A very different factor pattern was detected for this study sample than the factor pattern retained at the time when the 19 FII was standardised. Fifteen fields of interest were identified in 1969 by Alberts in his DPhil thesis, but neither factor scores, nor the sample size were reported in 1997 in the manual by Fouche and Alberts.

Six constructs (factors) were retained when working with this study sample. These six factors explained 66% of the variation in the data. Communalities varied from 0.37 to 0.81.

These results for the exploratory factor analysis, which were very different to those of the Fouche and Alberts (1977) study, were obtained even though all the conditions concerning the MSA, sample size, and skewness were more than satisfactorily met.

5.7.4. Evaluation

This study revealed that all the subtests of the 19 FII on the study sample are reliable with the implication that the 19 subtests' predictive validity can also be investigated in the future.

The scores of the 19 subtests can also be used to try to predict academic success by using logistic regression or a discriminant analysis.

Keeping in mind what construct validity means, namely that a test must measure what it is supposed to measure, the construct validity of the 19 FII for this study sample is questionable. Six indefinable constructs were retained with barely any resemblance to the fifteen retained at the time of standardisation. This means that it would not be wise to interpret an individual's score for a specific type of interest and use it to guide students to a future career. It would also not be fair to select an applicant for a specific field of study on the grounds of the results of this questionnaire's fields of interests.

Chapter 6

6. Predictors of Academic Success

This chapter addresses the third and fourth research questions namely:

3. Which of the available reliable and construct valid predictors are the best at predicting academic success for BCom, BPharm, BA, and BSc students, respectively?
4. Are there valid models which can adequately predict academic success for each of the BCom, BPharm, BA, and BSc students?

6.1. Introduction

As stated in Chapter 4, three statistical techniques were used in this study to answer these questions: CHAID, logistic regression, and predictive discriminant analysis. The results are discussed in the next sections and evaluations are made.

Variables

The outcomes of the first and second research questions of this study were used to decide which of the available subtests of psychometric tests were reliable and construct valid to be used as independent variables to enter the predictive models.

The following was reported in Chapter 5: All 10 subtests of the **SAT 78** were found reliable for the Study Sample (2003-2007) and the decision was made to use these 10 reliable subtests as possible independent variables for model fitting. *IQ* was calculated because all the subtests in its formula were reliable. Another reason for entering *IQ* is that *IQ* has high status in the educational environment in South Africa. By experimenting, *IQ* was then used instead of the separate subtests to enter the model so that its contribution to the models could be evaluated. The constructs obtained at the time of standardisation of the **SAT 78** in 1978 were very different from those obtained for the Study Sample (2003-2007) and were thus not used as independent variables to be entered in the models.

The 4 subtests as well as the constructs *Verbal, Nonverbal* and *Total* of the **GSAT** were reliable for the Study Sample (2003-2007). One construct, namely *Total,* was retained after an exploratory factor analysis was done, but the *Verbal* and *Non-verbal* constructs could not be separated by means of the factor analysis. In light of this, the decision was made to use the 4 reliable subtests of the GSAT as independent variables for model fitting. By

experimenting, the construct *Total* instead of the 4 subtests was also used to enter the model so that its contribution to the models could be evaluated.

The seven subtests of the **SSHA** were reliable for the Study Sample (2003-2007), but to prevent singularity (for example the sum of *Delay Avoidance* and *Work Methods* yields the subtests *Study Habits* - see Section 2.5.3.3) only *Delay Avoidance*, *Work Methods*, *Teacher Approval*, and *Education Acceptance* were used as variables to enter a model fitting procedure. When performing an exploratory factor analysis, one factor was retained, namely *Study Orientation.* By experimenting, *Study Orientation* instead of the four variables named above was also entered in the model to observe the effect on the model.

All the subtests of the **PHSF** were reliable for the Study Sample (2003-2007), but the constructs retained in this study were very different from those in 1983. The decision was made to enter all the variables except the *Desirability Scale* (see Section 2.5.4.3) of the PHSF as independent variables into the models.

The subtests of the **19FII** were reliable except *Work-Hobby* and *Active-Passive.* The constructs retained for the Study Sample (2003-2007) were very different from those obtained in 1983. Thus only the reliable subtests of the 19FII were used for model fitting.

6.2. Bachelor of Commerce

6.2.1. Study Sample

This sample consists of 385 students of which 54.55% were academic successes (66.19% were women and 33.81% were men). The academic failures were 45.45% (53.14% were women and 46.86% were men). However, due to the methods of handling missing data only 378 observations were used by the stepwise logistic regression procedure of which 54.50% were academic successes and 45.50% were academic failures. For similar reasons, in the stepwise discriminant analysis only 383 observations were used, 54.83% were academic successes and 45.17% were academic failures. The stepwise logistic regression analysis procedure in SAS uses only those students for whom there are no missing data for any variables, while the stepwise discriminant analysis procedure on the other hand demands only no missing data for the selected variables. The percentage of white students was 94.81%. The mean of this group's *Mscore* was 25.40 and it ranges from 12 – 36.

6.2.2. Variables

The dependent variable used in the CHAID procedure was the graduation status and the independent variables were variables 1 to 48 in Table 4.4, that is *Race*, *Gender*, *Year Group*, and *Mscore* and the reliable subtests of the SAT 78, SSHA, PHSF and the 19FII discussed in Section 6.1. Thus, 48 independent variables were entered into the CHAID procedure.

Multicollinearity

Multicollinearity was not present because when all the predictor variables were correlated there were no very high correlations. The highest correlation was 0.72 and the highest variance inflation factor (VIF) 4.06, which was under 5 (see Section 3.2.4).

Criteria

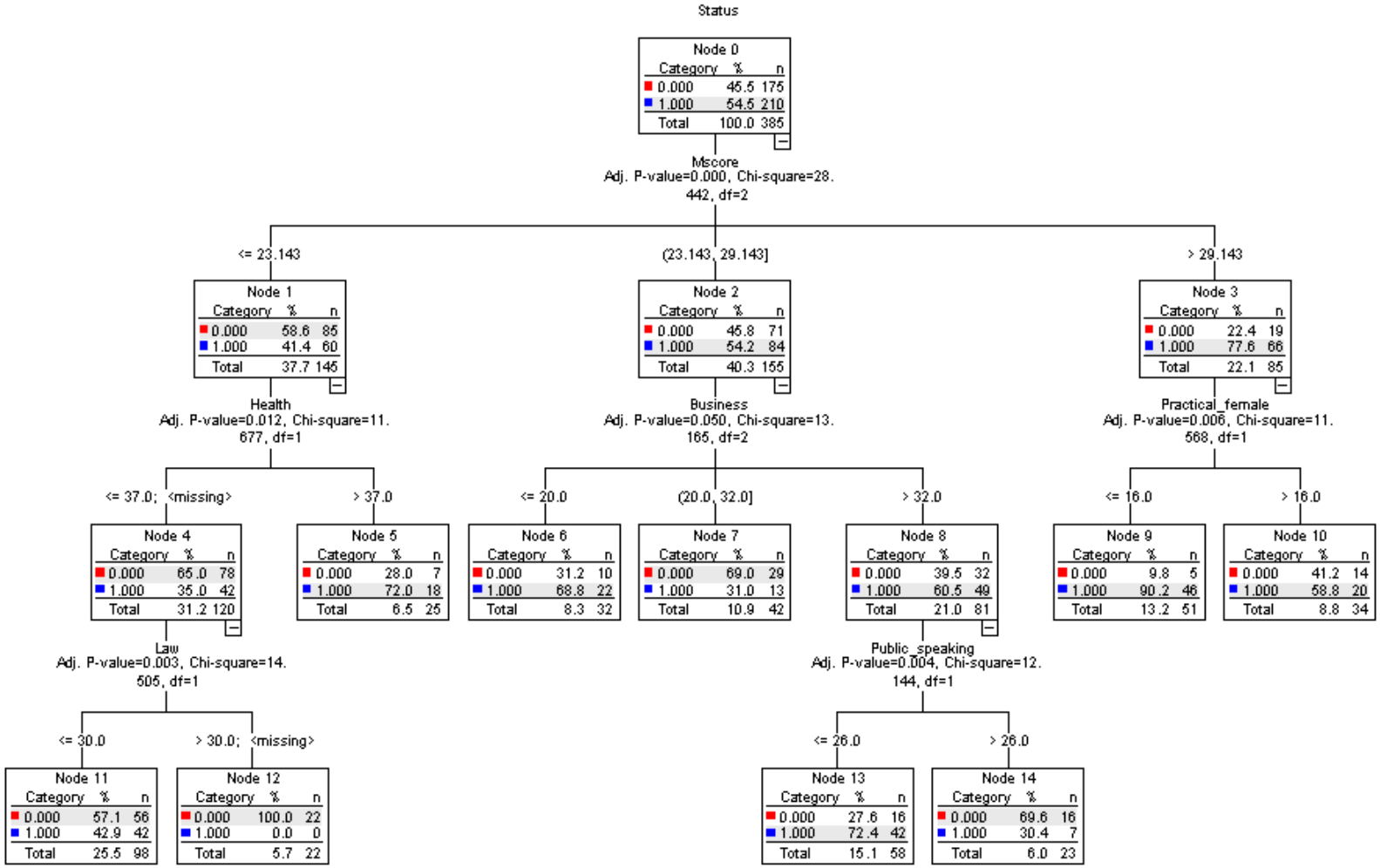
A p-level of 0.10 for splitting nodes and a p-level of 0.15 for merging categories were used in the CHAID analysis. The criteria for logistic regression and predictive discriminant analysis for a predictor to be entered in the model was a p-level of 0.10 and for a predictor to stay was 0.15.

6.2.3. CHAID: Results and Discussion

The CHAID analysis yielded the decision tree displayed in Figure 6.1.

The CHAID analysis revealed that *Mscore, Health, Business, Practical-female*, *Law, and Public Speaking* were the significant predictors of academic success for the BCom students. It also indicated that *Race*, *Gender*, and *Year Group* were not significant predictors and did not interact with any of the other significant predictor variables.

Figure 6.1 CHAID Tree Diagram for the BCom Group



6.2.4. Stepwise Logistic Regression: Results and Discussion

The results of the stepwise logistic regression analysis are given in Table 6.1. This result was obtained by entering the variables as described in Section 6.2.2 except Race (see section 4.7.3.1). By entering IQ and Study Orientation as described in Section 6.1 instead of the variables separately, the same set of variables was selected. The p-values in Table 6.1 are those obtained for the stepwise logistic regression analysis for the case of random sampling. The reciprocal of the odds ratio in cases where the odds ratio was less than one were computed for easier interpretation. Such variables in the context of all other variables in the model had a negative effect on academic success as can also be seen from by the sign of the estimate.

Table 6.1 Descriptive measures of the stepwise logistic analysis: BCom group

Variable	Estimate	Standard Error	Chi-square	P-value	Effect size	Adjusted Odds Ratio	1/Adjusted Odds Ratio	Correlation coefficient with logit
Intercept	-3.44	0.71	23.32	<0.0001	0.25			
Self-control	0.04	0.02	2.75	0.10	0.09	1.21		0.32
Social Work	0.03	0.01	5.21	0.02	0.12	1.32		0.04
Public Speaking	-0.03	0.01	4.55	0.03	0.11	0.75	1.33	-0.41
Law	-0.02	0.01	3.75	0.05	0.10	0.77	1.30	-0.42
Disguised Words	-0.04	0.02	3.19	0.07	0.09	0.81	1.23	0.01
Figure Series	0.02	0.01	4.47	0.03	0.11	1.28		0.25
Mscore	0.16	0.03	35.76	<.0001	0.31	2.20		0.72

$n = 378$ $n_1 = 206$ $n_0 = 172$

Seven predictor variables were selected with the stepwise selection method, namely *Self-control, Social Work, Public Speaking, Law, Disguised Words, Figure Series,* and *Mscore.* The number of predictors p in the model must meet the requirement that $p+1 < \min(n_1, n_0)/10 = 17.2$. For this model it is met because $p+1 = 8$ (see Sections 3.2.2.9 and 4.7.3.1).

Linearity with the logit

To interpret the odds ratio for continuous variables correctly, the requirement is that the relationship between the specific variable and the logit must be linear. The correlation coefficient (that is the effect size, see Section 3.3.2) for *Self-control, Public Speaking, Law,* and *Figure Series* with the logit were all approximately of a medium effect (see Table 6.1). That means that a linear relationship between these separate variables and the logit could be

observed by the naked eye. The correlation coefficient of *Mscore* with the logit is 0.72 which is a practically significant linear relationship. *Disguised Words* and *Social Work* are not linearly related to the logit and no transformation could be found which made the relationship for these two variables linear with the logit.

Mscore is the predictor with the highest odds ratio, namely 2.20, and is according to Section 3.3.4 the only predictor which is also practically significant. An odds ratio of 2.20 indicates that for every increase of 4.85 (one standard deviation, see Section 4.7.3.2) in *Mscore* the chance for completing a BCom degree in the prescribed time increases by a factor 2.2, if the other predictors in the model are held constant.

Table 6.2 gives the results of the Hosmer and Lemeshow goodness-of-fit test.

Table 6.2 Hosmer and Lemeshow goodness-of-fit test: BCom group

		Status = 1		*Status = 0*	
Group	Total	Observed	Expected	Observed	Expected
1	38	8	7.47	30	30.53
2	38	17	11.97	21	26.03
3	38	11	15.08	27	22.92
4	38	19	17.9	19	20.1
5	38	17	20.04	21	17.96
6	39	24	22.82	15	16.18
7	38	23	24.45	15	13.55
8	38	26	26.76	12	11.24
9	38	31	29.68	7	8.32
10	35	30	29.84	5	5.16

n = 378 chi-square = 6.80 p-value = 0.59 (in case of random sampling)
w = 0.13 df = 8.

The chi-square value of the Hosmer and Lemeshow goodness-of-fit test was 6.80 and the degrees of freedom 8. The effect size is then w = 0.13, which is a small effect size (defined as of no practical significance) and thus means that the fit of the model is good (see Section 3.3.3). The expected frequencies in all cells are above 5, which make the conclusion that the model fits valid (Hosmer & Lemeshow, 2000).

Area under ROC curve

The area under the ROC curve was found to be 0.75, which is considered acceptable discrimination (see 3.2.2.9).

Cross validation

The leave-one-out principle was used for cross validation of the logistic regression procedure. That is, dropping the data of one subject at a time and then re-estimating the parameter estimates to classify that subject. A cut-off point of 0.5 was used for the classification. Table 6.3 gives the classification according to the logistic regression analysis.

Table 6.3 Classification table for the stepwise logistic regression: BCom group

		Predicted Status		
		0	**1**	**Total**
Actual Status	**0**	96	76	**172**
		55.81	44.19	**100**
	1	50	156	**206**
		24.27	75.73	**100**
Total		**146**	**232**	**378**
		38.62	**61.38**	**100**

Improvement over chance

The effect size index *I* for improvement over chance was calculated from Table 6.3. The observed hit rate (H_o) was (96+156)/378 = 0.67 and the expected hit rate (H_e) was 0.50 with n = 378. By using Equation 3.46, *I* was calculated for the logistic regression model and found to be equal to 0.34, which is an almost practically significant improvement over chance.

Best subset selection

The seven predictors selected as the best subset of seven predictors with the logistic regression procedure were *Self-control, Social Work, Public Speaking, Law, Disguised Words, Figure Series, and Mscore.* The highest global score test chi-square value *C* for a selection of five predictors was 58.94 (see Section 3.2.2.8).

6.2.5. Stepwise Predictive Discriminant Analysis: Results and Discussion

The multivariate distribution of the variables used for predictive discriminant analysis was not multivariate normal, but according to McLachlan (2004) linear discriminant analysis is fairly robust against departure from this assumption. To determine whether a linear classification rule or a quadratic rule had to be used, a chi-square test was done to test for equal covariance matrices of the two groups. The chi-square value was 29.18, the degrees of freedom 28 and the p-value = 0.40 (in the case of random sampling). By calculating the effect size *w* (see Section 3.3.3), the value *w* = 0.30, which is not practically significant meaning that a pooled covariance matrix and thus a linear classification rule could be used. Table 6.4 gives the coefficients of the linear discriminant functions for the two status groups. The prior probabilities were chosen as 0.45 (45%) for the academic unsuccessful group and 0.55 (55%) for the academic successful group, in accordance with the percentage of failures and successes in the sample.

Table 6.4 Linear Discriminant Function for status: BCom group

Variable	Status = 0	Status = 1
Constant	-35.15	-39.36
Mscore	1.11	1.26
Public Speaking	0.07	0.55
Social Work	0.18	0.20
Law	0.67	0.05
Figure Series	0.02	0.05
Disguised Words	0.26	0.22
Self-control	1.23	1.27
Priors	**0.45**	**0.55**

$n = 383$ $n_1 = 210$ $n_0 = 173$

Seven predictors were selected with the stepwise selection procedure: *Self-control, Social Work, Public Speaking, Law, Disguised Words, Figure Series,* and *Mscore.*

Cross validation

The leave-one-out principle was used for cross validation of the discriminant analysis procedure. An observation was classified by calculating its value for both linear discriminant functions (based on the data without that observation) and then classified into the group for

which the value of the function was the highest. According to Section 3.2.3.2 for this method to be valid the requirement is that $n_j > 3p$ where p is the number of predictors and $n_j = \min(n_0, n_1)$. The requirement was met, because, $p = 7$ and $3p = 21$. An observation was classified by calculating its value for both linear discriminant functions (based on the data without that observation) and then classified into the group for which the value of the function was the highest. Table 6.5 gives the classification according to the predictive discriminant analysis.

Table 6.5 Classification table for the stepwise predictive discriminant analysis: BCom group

		Predicted Status		
Actual Status		**0**	**1**	**Total**
	0	92	81	**173**
		53.18	46.82	**100**
	1	52	158	**210**
		24.76	75.24	**100**
Total		**144**	**239**	**383**
		37.60	**62.40**	**100**

Improvement over chance

For the discriminant analysis model the observed hit rate (H_o) was (92+158)/383 = 0.65 and the expected hit rate (H_e) was (0.38x173 + 0.62x210)/383 = 0.50, and n = 383 (from Table 6.5). When *I* was calculated for this model it was found to be equal 0.30, which is a medium effect size for over chance (see Section 3.3.5).

6.2.6. Evaluation

The same seven predictors that were selected by the stepwise logistic regression, namely *Self-control, Social Work, Public Speaking, Law, Disguised Words, Figure Series,* and *Mscore,* were also selected as the best subset of seven predictors in logistic regression as well as the stepwise discriminant analysis for the BCom group. However, a very different selection of significant predictors was selected by the CHAID procedure. *Mscore, Health,*

Business, Practical-female, *Law,* and *Public Speaking*. *Mscore, Law,* and *Public Speaking* were thus selected by all four procedures (that is stepwise logistic regression, best subset of seven predictors in logistic regression, discriminant analysis and CHAID).

The sensitivity of the model fitted by logistic regression is 75.73% (156/206) and the specificity is 55.81% (96/172), as was calculated from Table 6.3. The sensitivity of the model fitted by predictive discriminant analysis is 75.24% (158/210) and the specificity is 53.18% (92/173), as was calculated from Table 6.5. When comparing the students who were misclassified by logistic regression, with those who were misclassified by predictive discriminant analysis, there were four students who were misclassified by discriminant analysis, but correctly classified by logistic regression. Those four students were wrongly classified by discriminant analysis as to be successes, while they were actually failures. This result is confirmed by the fact that the specificity of the discriminant analysis is approximately two percent lower than that of the specificity of logistic regression.

6.3. Bachelor of Pharmacy

6.3.1. Study Sample

This sample consists of 267 students of which 77.90% were academic successes (84.62% were women and 15.38% were men). The academic failures were 22.10% (25.42% were women and 74.58% were men). However, due to the methods of handling missing data only 255 observations were used by the logistic regression procedure of which 77.65% were academic successes and 22.35% were academic failures. For similar reasons, in the discriminant analysis only 260 observations were used of which 77.69% were academic successes and 22.31% were academic failures (see Section 6.2.1). The percentage of white students was 96.25%. The mean of this group's *Mscore* was 29.72 and it ranges from 16 – 36.

6.3.2. Variables

The dependent variable used was the graduation status and the independent variables were variables 15 to 55 in Table 4.4, that is the reliable subtests of the GSAT, SSHA, PHSF and the 19FII discussed in Section 6.1 as well as the *Chemistry*, *Mathematics* and *Physics* marks obtained (see Section 4.5). *Race*, *Gender*, *Year Group*, and *Mscore* were also entered as independent variables. The pharmacy students completed the GSAT instead of the SAT 78. Thus, 45 independent variables were entered into the CHAID procedure.

Multicollinearity

There was a high correlation of 0.84 between *Teacher Approval* and *Education Acceptance* and the VIF of *Education Acceptance* was more than 5. As a result of the fact that the VIF of *Education Acceptance* was more then 5, the decision was made to omit *Education Acceptance* from the list of variables that may be used in the model.

After omitting *Education Acceptance* multicollinearity was not found to be present, because none of the correlation coefficients between the remaining predictor variables were above 0.71 and the highest VIF was 3.66, which was under 5 (see Section 3.2.4).

Criteria

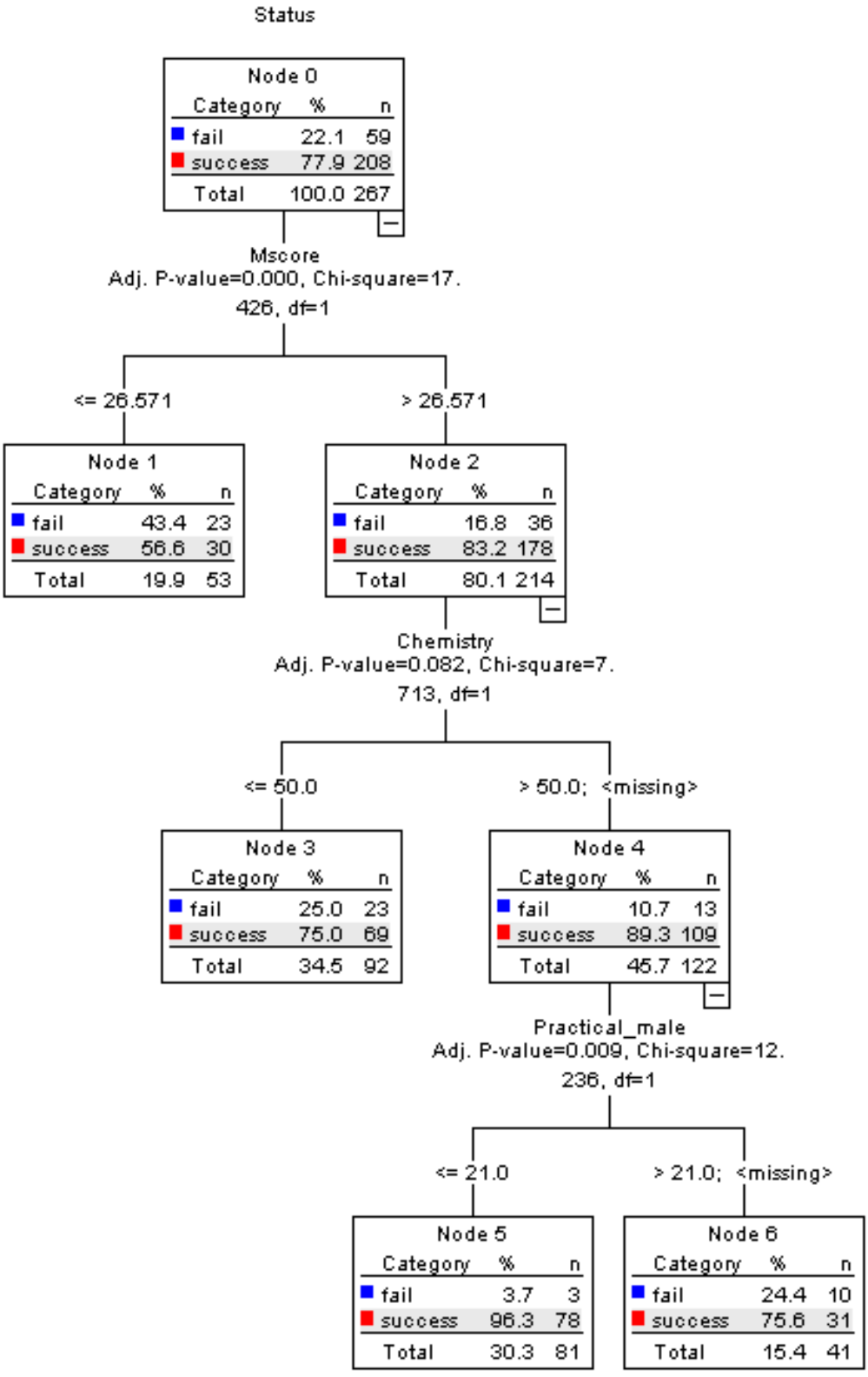
A level of 0.15 for splitting nodes and a level of 0.20 for merging categories were used in the CHAID analysis. The criteria for logistic regression and predictive discriminant analysis for a predictor to be entered in the model was 0.15 and for a predictor to stay was 0.20.

6.3.3. CHAID: Results and Discussion

The CHAID analysis yielded the decision tree displayed in Figure6.2.

The CHAID analysis revealed that *Mscore, Chemistry,* and *Practical-male*, were the significant predictors of academic success for the BPharm students. It also indicated that *Race*, *Gender*, and *Year Group* were not significant and did not interact with any of the other significant predictor variables.

Figure 6.2 CHAID Tree Diagram for the BPharm Group



6.3.4. Stepwise Logistic Regression: Results and Discussion

The results of the stepwise logistic regression analysis are given in Table 6.6. This was obtained by entering the variables as described in Section 6.3.2 except *Race* (see section 4.7.3.1). By entering *Verbal* and *Study Orientation* as described in Section 6.1 instead of the variables separately, the same variables were selected. In Table 6.6 descriptive measures for the stepwise logistic regression are given. The p-values in Table 6.6 are those obtained for the logistic analysis for the case for random sampling. The reciprocal of the odds ratio in cases where the odds ratio was less than one were computed for easier interpretation. Such variables in the context of all other variables in the model had a negative effect on academic success as can also be seen from the sign of the estimate.

Table 6.6 Descriptive measures of the stepwise logistic analysis: BPharm group

Variable	*Estimate*	*Standard Error*	*Chi-square*	*P-value*	*Effect size*	*Adjusted Odds Ratio*	*1/Adjusted Odds Ratio*	*Correlation coefficient with logit*
Intercept	-6.41	1.76	13.20	0.00	0.22			
Family Influences	0.07	0.02	7.23	0.01	0.17	1.53		0.54
Science	0.03	0.02	3.23	0.07	0.11	1.32		0.42
Public Speaking	0.03	0.02	2.31	0.13	0.10	1.34		-0.01
Law	-0.04	0.02	5.76	0.02	0.15	0.64	1.56	-0.37
Mscore	0.16	0.05	10.68	0.00	0.20	1.73		0.64

$n = 255$ $n_1 = 198$ $n_0 = 57$

Five predictor variables were selected with the stepwise selection method, namely *Family Influences*, *Science, Public Speaking, Law,* and *Mscore.* The number of predictor variables p in the model must meet the requirement that $p+1 \leq \min(n_1, n_0)/10 = 5.7$. For this model it is met because $p+1 = 6$, which is very close to 5.7 and the rule is not that strict (Hosmer & Lemeshow, 2000) (see Section 3.2.2.9).

Linearity with the logit

To interpret the odds ratio for continuous variables correctly, the assumption is that the relationship between the specific variable and the logit must be linear. The absolute value of the correlation coefficient (that is the effect size, see Section 3.3.2) for *Science* and *Law,* with the logit were both approximately of a medium effect see Table 6.6. That means that a linear relationship between these separate variables and the logit could be observed by the naked eye. The correlation coefficient of *Mscore* and *Family Influences* with the logit is 0.54 and

0.64 respectively which is a practical significant linear relationship. *Public Speaking* is not linearly related to the logit and no transformation could be found which made the relationship for this variable linear with the logit.

Mscore is the predictor with the highest odds ratio namely 1.73 but is according to Section 3.3.4 not practically significant. However an odds ratio of 1.73 indicates that for every increase of 3.45 (one standard deviation, see Section 4.7.3.1) in *Mscore* the chance for completing a BPharm degree in the prescribed time increases by a factor 1.73, if the other predictors in the model are held constant.

Table 6.7 gives the results of the Hosmer and Lemeshow goodness-of-fit test.

Table 6.7 Hosmer and Lemeshow goodness-of-fit test: BPharm group

		Status=1		*Status=0*	
Group	**Total**	**Observed**	**Expected**	**Observed**	**Expected**
1	26	12	12.16	14	13.84
2	26	15	16.49	11	9.51
3	26	16	18.27	10	7.73
4	26	22	19.76	4	6.24
5	26	24	20.72	2	5.28
6	26	20	21.59	6	4.41
7	26	23	22.29	3	3.71
8	26	24	23.02	2	2.98
9	26	23	23.75	3	2.25
10	21	19	19.94	2	1.06

n = 255 chi-square = 7.28 p-value = 0.51 (in case of random sampling)
w = 0.17 df = 8.

The chi-square value of the Hosmer and Lemeshow goodness-of-fit test was 7.28 and the degrees of freedom 8. The effect size w = 0.17, which is a small effect size (defined as of no practical significance) and thus means that the fit of the model is good (see Section 3.3.3). Five of the cells have expected frequencies less than 5, which makes the conclusion that the model fits less valid (Hosmer & Lemeshow, 2000).

Area under ROC curve

The area under the ROC curve was found to be 0.72, which is considered acceptable discrimination (see Section 3.2.2.9).

Cross validation

The leave-one-out principle was used for cross validation of the stepwise logistic regression procedure. That is, dropping the data of one subject at a time and then re-estimating the parameter estimates to classify that subject. A cut-off point of 0.5 was used for the classification. Table 6.8 gives the classification according to the logistic regression analysis.

Table 6.8 Classification table for the stepwise logistic regression: BPharm group

		Predicted Status		
		0	**1**	**Total**
Actual Status	**0**	6	51	**57**
		10.53	89.47	**100**
	1	7	191	**198**
		3.45	96.46	**100**
Total		**13**	**242**	**255**
		5.09	**94.90**	**100**

Improvement over chance

The effect size index *I* for improvement over chance was calculated from Table 6.8. The observed hit rate (H_o) was (6+191)/255 = 0.77 and the expected hit rate (H_e) was 0.65 with n = 255. By using Equation 3.46, *I* was calculated for the logistic regression model and found to be equal to 0.34, which is a practical significant improvement over chance, suggesting a valid model.

Best subset of five selection

The five predictors selected as the best subset with the logistic regression procedure were *Family Influences*, *Science*, *Public Speaking*, *Law*, and *Mscore*. The highest global score test chi-square value *C* for a selection of five predictors was 27.99 (see Section 3.2.2.8).

6.3.5. Stepwise Predictive Discriminant Analysis: Results and Discussion

To determine whether a linear classification rule or a quadratic rule had to be used, a chi-square test was used to test for equal covariance matrices of the two groups. The chi-square value was 18.14, the degrees of freedom 15 and the p-value = 0.26 (in the case of random

sampling). By calculating the effect size *w* (see Section 3.3.3), the value *w* = 0.26, which is not practically significant meaning that a pooled covariance matrix and thus a linear classification rule could be used. Table 6.9 gives the coefficients of the linear discriminant functions for the two status groups. The prior probabilities were chosen as 0.22 (22%) for the academic unsuccessful group and 0.78 (78%) for the academic successful group, in accordance with the percentage of failures and successes in the sample.

Table 6.9 Linear Discriminant Function for status: BPharm group

Variable	Status = 0	Status = 1
Constant	-55.99	-62.68
Mscore	2.51	2.68
Family Influences	0.86	0.92
Law	-0.01	-0.05
Science	0.29	0.33
Public Speaking	0.23	0.25
Priors	**0.22**	**0.78**

$n = 260$ $n_1 = 202$ $n_0 = 58$

Five predictors were selected by the stepwise discriminant analysis, namely *Family Influences*, *Science*, *Public Speaking*, *Law*, and *Mscore*.

Cross validation

The leave-one-out principle was used for cross validation of the discriminant analysis procedure. An observation was classified by calculating its value for both linear discriminant functions (based on the data without that observation) and then classified into the group for which the value of the function was the highest. According to Section 3.2.3.2 for this method to be valid the requirement is that $n_j > 3p$ where p is the number of predictors and $n_j = \min(n_0, n_1) = 58$. The requirement was met, because, $p = 5$ and $3p = 15$, which is smaller than 58. Table 6.10 gives the classification according to the predictive discriminant analysis.

Table 6.10 Classification table for stepwise predictive discriminant analysis: BPharm group

		Predicted Status		
Actual Status		**0**	**1**	**Total**
	0	5	53	**58**
		8.62	91.38	**100**
	1	9	193	**202**
		4.46	95.54	**100**
Total		**14**	**246**	**260**
		5.39	**94.61**	**100**

Improvement over chance

For the discriminant analysis model the observed hit rate (H_o) was (5+193)/260 = 0.76 and the expected hit rate (H_e) was 0.65 with n = 260 (from Table 6.10). When *I* was calculated for this model, it was found to be equal to 0.31, which is a medium to large effect size of improvement over chance (see Section 3.3.5).

6.3.6. Evaluation

The same five predictors that were selected by stepwise logistic regression, namely *Family Influences*, *Science*, *Public Speaking*, *Law*, and *Mscore,* were also selected as the best subset of five predictors in logistic regression as well as by the stepwise discriminant analysis for the BPharm group. However, a very different selection of significant predictors was selected by the CHAID procedure, namely *Mscore, Chemistry,* and *Practical-male*. *Mscore* is thus the only predictor that was selected by all four procedures (that is stepwise logistic regression, best subset of five predictors in logistic regression, discriminant analysis and CHAID).

The sensitivity of the model fitted by logistic regression is 96.46% (191/198) and the specificity is 10.53% (6/57), as can be seen in Table 6.8. The sensitivity of the model fitted by predictive discriminant analysis is 95.54% (193/202) and the specificity is 8.62 (5/58), as can be seen in Table 6.10. When comparing the students who were misclassified by logistic regression, with those who were misclassified by predictive discriminant analysis there were

three students who were misclassified by discriminant analysis, but correctly classified by logistic regression. Two of those three students were wrongly classified by discriminant analysis as failures, while they were actually successes and one failure was wrongly classified as a success. This result is confirmed by the fact that the sensitivity and specificity of the discriminant analysis are lower than the sensitivity and specificity of logistic regression respectively.

6.4. Bachelor of Arts

6.4.1. Study Sample

This sample consists of 177 students of which 36.16% were academic successes (84.38% were women and 15.62% were men). The academic failures were 63.84% (78.76% were women and 21.24% were men). However, due to the methods of handling missing data only 172 observations were used by the logistic regression procedure of which 37.21% were academic successes and 62.79% were academic failures. For similar reasons, in the discriminant analysis only 174 observations were used, 36.78% were academic successes and 63.22% were academic failures (see Section 6.2.1). The percentage of white students was 89.83%. The mean *Mscore* of this group is 22.36 and it ranges from 10 – 36.

6.4.2. Variables

The dependent variable used was the graduation status and the independent variables were variables 1 to 48 in Table 4.4, that is *Race*, *Gender*, *Year Group*, and *Mscore* and the reliable subtests of the SAT 78, SSHA, PHSF and the 19FII discussed in Section 6.1. Thus, 48 independent variables were entered into the CHAID procedure.

Multicollinearity

Multicollinearity was not present because when all the predictor variables were correlated there were no very high correlations. The highest correlation was 0.73 between *Work-Methods* and *Delay Avoidance* (see Section 3.2.4). The VIF of *Delay Avoidance* was 5.06, which was slightly above 5. As a result of this fact the decision was made to let both these variables enter the stepwise logistic procedure so as to give both of them the chance to be selected as predictor variable in the model building stage of the analysis.

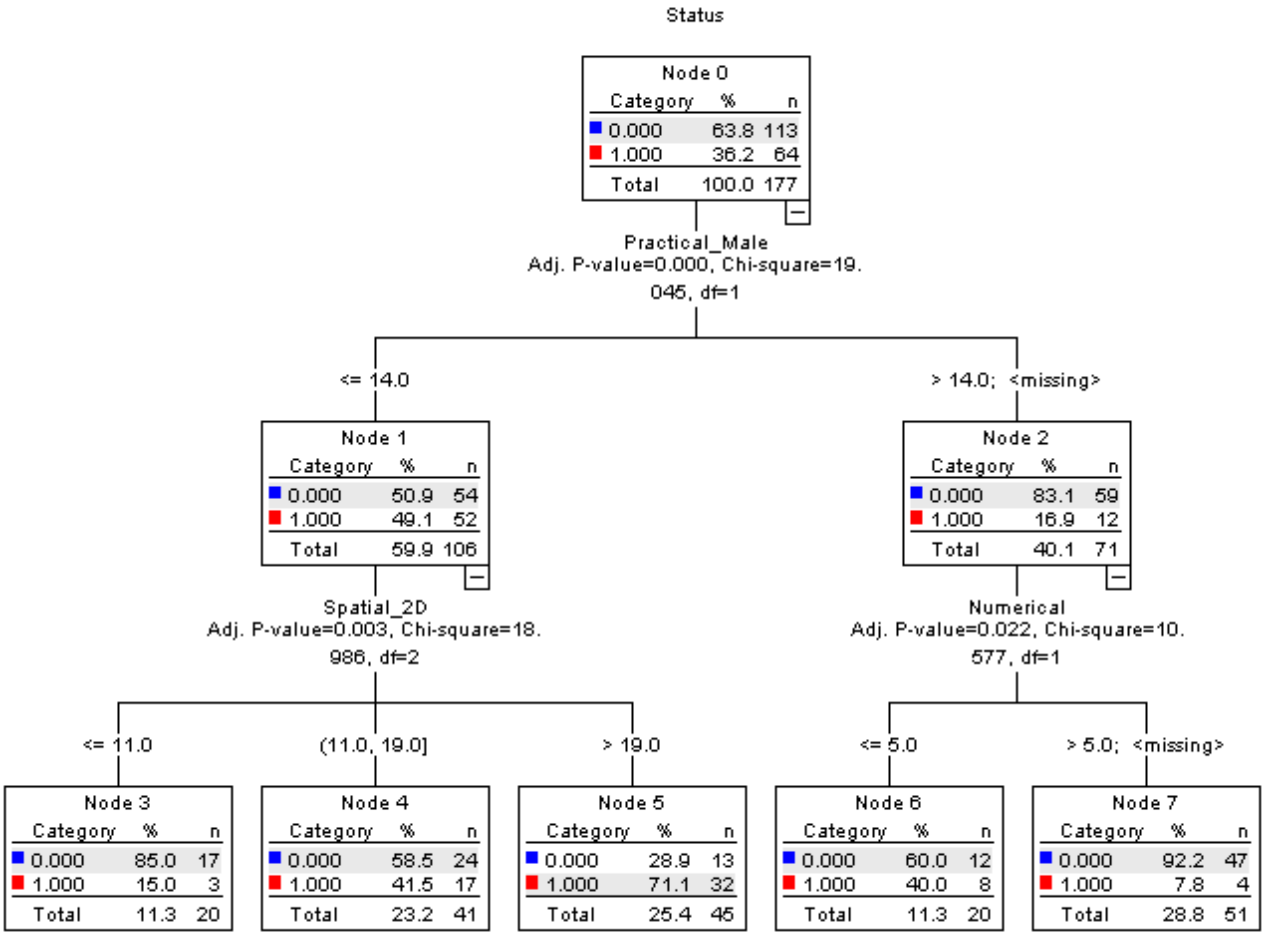
Criteria

A p-level of 0.05 for splitting nodes and a p-level of 0.10 for merging categories were used in the CHAID analysis. The criteria for stepwise logistic regression and stepwise predictive discriminant analysis for a predictor to be entered in the model was a p-level of 0.05 and for a predictor to stay was 0.10.

6.4.3. CHAID: Results and Discussion

The CHAID analysis yielded the following decision tree.

Figure 6.3 CHAID Tree Diagram for the BA Group



The CHAID analysis revealed that *Practical-Male*, *Spatial 2D*, and *Numerical* were the significant predictors of academic success for the BA students. It also indicated that *Race*, *Gender*, and *Year Group* were not significant predictors and did not interact with any of the other significant predictor variables.

6.4.4. Stepwise Logistic Regression: Results and Discussion

The results of the stepwise logistic regression analysis are given in Table 6.11. This was obtained by entering the variables as described in Section 6.4.2 except *Race* (see section 4.7.3.1). By entering *Study Orientation* as described in Section 6.1 instead of the variables separately, the same variables were selected. In Table 6.11 descriptive measures for the stepwise logistic regression are given. The p-values in Table 6.11 are those obtained for the logistic analysis for the case of random sampling. The reciprocal of the odds ratio in the case where the odds ratio was less than one were computed for easier interpretation. Such variables in the context of all other variables in the model had a negative effect on academic success as can also be seen from the sign of the estimate.

Table 6.11 Descriptive measures of the logistic analysis: BA group

Variable	Estimate	Standard Error	Chi-square	P-value	Effect size	Adjusted Odds Ratio	1/Adjusted Odds Ratio	Correlation coefficient with logit
Intercept	-9.70	2.03	22.71	<0.0001	0.36			
Self-control	0.12	0.04	9.20	0.00	0.23	1.82		0.41
Nature	-0.05	0.02	5.81	0.02	0.18	0.57	1.75	-0.60
Comparison	0.14	0.06	5.89	0.02	0.19	1.63		0.53
Spatial 2D	0.08	0.03	6.50	0.01	0.19	1.69		0.56
Mscore	0.09	0.04	5.75	0.02	0.18	1.59		0.53

$n = 172$ $n_1 = 64$ $n_0 = 108$

Five predictor variables were selected with the stepwise selection method, namely *Self-control, Nature, Comparison, Spatial 2D,* and *Mscore.* The number of parameters p in the model must meet the requirement that $p+1 \leq \min(n_1, n_0)/10 = 6.4$. For this model it is met because $p+1 = 6$ (see Section 3.2.2.9).

Linearity with the logit

To interpret the odds ratio for continuous variables correctly, the requirement is that the relationship between the specific variable and the logit must be linear. The correlation coefficient (that is the effect size, see Section 3.3.2) for *Self-control* with the logit was approximately of a medium effect (see Table 6.11). That means that a linear relationship between this variable and the logit could be observed by the naked eye. Each of the absolute values of the correlation coefficients of *Nature, Comparison, Spatial 2D,* and *Mscore* with the logit is above 0.5, and thus all of them have practically significant linear relationships with the logit.

Self-control is the predictor with the highest odds ratio, namely 1.82, and is according to Section 3.3.4 not practically significant. An odds ratio of 1.82 indicates that for every increase of 4.85 (one standard deviation) in *Self-control* the chance for completing a BA degree in the prescribed time increases by a factor 1.82, if the other predictors in the model are held constant.

Table 6.12 gives the results of the Hosmer and Lemeshow goodness-of-fit test.

Table 6.12 Hosmer and Lemeshow goodness-of-fit test: BA group

		Status = 1		*Status = 0*	
Group	Total	Observed	Expected	Observed	Expected
1	17	1	0.72	16	16.28
2	17	1	1.36	16	15.64
3	17	2	2.61	15	14.39
4	17	4	3.88	13	13.12
5	17	6	5.29	11	11.71
6	17	4	6.57	13	10.43
7	17	12	7.92	5	9.07
8	17	9	9.17	8	7.83
9	17	10	11.25	7	5.57
10	19	15	15.21	4	3.79

n = 172 chi-square = 6.51 p-value = 0.59 (in case of random sampling)
w = 0.19 df = 8.

The chi-square value of the Hosmer and Lemeshow goodness-of-fit test was 6.51 and the degrees of freedom 8. The effect size then is w=0.19 which is a small effect size (defined as of no practical significance) and thus means that the fit of the model is good (see Section 3.3.3). Five of the cells have expected frequencies less than 5, which makes the conclusion that the model fits, less valid (Hosmer & Lemeshow, 2000).

Area under ROC curve

The area under the ROC curve was found to be 0.75, which is considered acceptable discrimination (see Section 3.2.2.9).

Cross validation

The leave-one-out principle was used for cross validation of the logistic procedure. That is, dropping the data of one subject at a time and then re-estimating the parameter estimates to

classify that subject. A cut-off point of 0.5 was used for the classification. Table 6.13 gives the classification according to the logistic regression analysis.

Table 6.13 Classification table for the stepwise logistic regression: BA group

		Predicted Status		
		0	**1**	**Total**
Actual Status	**0**	91	17	**108**
		84.26	15.74	**100**
	1	30	34	**64**
		46.88	53.12	**100**
Total		**121**	**51**	**172**
		70.35	**29.65**	**100**

Improvement over chance

The effect size index *I* for improvement over chance was calculated from Table 6.13. The observed hit rate (H_o) was (91+34)/172 = 0.73 and the expected hit rate (H_e) was 0.53 with n = 172. By using Equation 3.46, *I* was calculated for the logistic regression model and found to be equal to 0.43, which is a practical significant improvement over chance.

Best subset selection

The five predictors selected as the best subset of five predictors with the logistic regression procedure were *Self-control, Practical-Male, Comparison, Spatial 2D,* and *Mscore*. The highest global score test chi-square value *C* for a selection of five predictors was 42.57 (see Section 3.2.2.8).

6.4.5. Stepwise Predictive Discriminant Analysis: Results and Discussion

To determine whether a linear classification rule or a quadratic rule had to be used, a chi-square test was used to test for equal covariance matrices of the two groups. The chi-square value was 21.27, the degrees of freedom 15 and the p-value = 0.13 (in the case of random sampling). By calculating the effect size *w* (see Section 3.3.3), the value *w* = 0.35, which is not practically significant meaning that a pooled covariance matrix and thus a linear classification rule could be used. Table 6.14 gives the coefficients of the linear discriminant functions for the two status groups. The prior probabilities were chosen as 0.64 (64%) for the academic unsuccessful group and 0.36 (36%) for the academic successful group, in accordance with the percentage of failures and successes in the sample.

Table 6.14 Linear Discriminant Function for status: BA group

Variable	Status = 0	Status = 1
Intercept	-50.95	-61.11
Self-control	0.66	0.75
Nature	0.26	0.22
Comparison	0.36	0.43
Spatial 2D	1.53	1.65
Mscore	2.02	2.18
Priors	**0.64**	**0.36**

n = 174 n_1 = 64 n_0 = 110

Five predictors were selected with the stepwise selection procedure, namely *Self-control, Nature, Comparison*, *Spatial 2D,* and *Mscore.*

Cross validation

The leave-one-out principle was used for cross validation of the discriminant analysis procedure. An observation was classified by calculating its value for both linear discriminant functions (based on the data without that observation) and then classified into the group for which the value of the function was the highest. According to Section 3.2.3.2 for a selection of five predictors $n_j > 3p$, as $n_j = 64$, $p = 5$, and $3p = 15$ and the requirement is thus met.

Table 6.15 gives the classification according to the predictive discriminant analysis.

Table 6.15 Classification table for stepwise predictive discriminant analysis: BA group

	Predicted Status		
Actual Status	**0**	**1**	**Total**
0	91	19	**110**
	82.73	17.27	**100**
1	31	33	**64**
	48.44	51.56	**100**
Total	**122**	**52**	**174**
	70.11	**29.89**	**100**

Improvement over chance

For the discriminant analysis model the observed hit rate (H_o) was (91+33)/174 = 0.73 and the expected hit rate (H_e) was 0.53 with n = 174 (from Table 6.15). When I was calculated for this model it was found to be equal to 0.38, which is also a practical significant improvement over chance (see Section 3.3.5).

6.4.6. Evaluation

The same five predictors that were selected by the stepwise logistic regression, namely *Self-control, Nature, Comparison*, *Spatial 2D,* and *Mscore*, were also selected as the best subset of five predictors in logistic regression as well as by the stepwise discriminant analysis for the BA group. The five predictors that were selected as the best subset selection in logistic regression were also similar to the logistic and discriminant analysis' stepwise selection procedures. However, a very different selection of significant predictors was selected by the CHAID procedure, namely *Practical-Male*, *Spatial 2D*, *Numerical*. *Spatial 2D* is thus the only predictor that was selected by all four the procedures.

The sensitivity of the model fitted by logistic regression is 53.12% (34/64) and the specificity is 84.26% (91/108), as can be seen from Table 6.13. The sensitivity of the model fitted by predictive discriminant analysis is 51.56% (33/64) and the specificity is 82.73% (91/110), as can be seen in Table 6.15. When comparing the students who were misclassified by logistic regression, with those who were misclassified by predictive discriminant analysis there were

two students who were misclassified by discriminant analysis, but correctly classified by logistic regression. One of these students was wrongly classified by discriminant analysis as a success, while this student was actually a failure while the other one was a success and was classified as a failure. This result is confirmed by the fact that the sensitivity as well as the specificity of the discriminant analysis is approximately one percent lower than the sensitivity and specificity of logistic regression. There was a slight difference in the number of observations used by the different procedures, which is the reason why both procedures can yield the same amount (91) of failures that was correctly classified. One of the 91 observations which discriminant analysis classified as a failure was not part of the observations used by logistic regression.

6.5. Bachelor of Science

6.5.1. Study Sample

This sample consists of 150 students of which 39.33% were academic successes (77.97% were women and 22.03% were men). The academic failures were 60.67% (60.44% were women and 39.56% were men). Of the 150 students 82 (54.67%) were women and 68 (45.33%) were men. However, due to the methods of handling missing data only 149 observations were used by the logistic regression procedure of which 39.60% were academic successes and 60.40% were academic failures. For similar reasons, in the discriminant analysis only 149 observations were used, 39.60% were academic successes and 60.40% were academic failures. The percentage of white students was 97.33%. The mean of this group's *Mscore* was 26.07 and it ranges from 17 – 36.

6.5.2. Variables

The dependent variable used was the graduation status and the independent variables were variables 2 to 48 in Table 4.4, that is *Race*, *Gender*, *Year Group*, and *Mscore* and the reliable subtests of the SAT 78, SSHA, PHSF and the 19FII, discussed in Section 6.1. Thus 48 independent variables were entered into the CHAID procedure.

Multicollinearity

There was a high correlation of 0.77 between *Delay Avoidance* and *Work Methods* and the VIF of *Work Methods* was more than five. The decision was made to omit *Work Methods* from the list of variables to be used in the model.

After omitting *Work Methods* multicollinearity was not found to be present, because when all the remaining predictor variable were correlated there were no correlations above 0.78 and the highest VIF was 4.50, which was under 5 (see Section 3.2.4).

Criteria

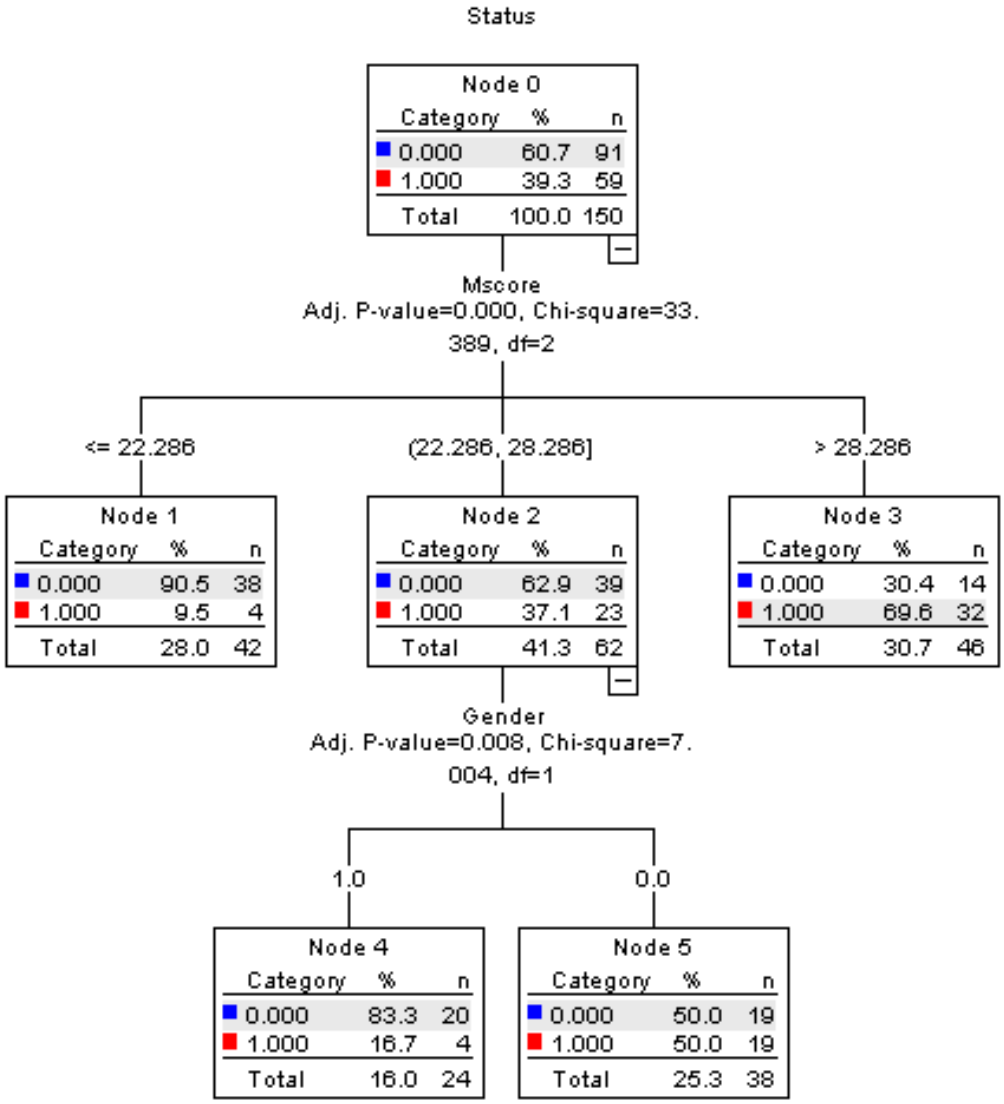
A p-level of 0.05 for splitting nodes and a p-level of 0.10 for merging categories were used in the CHAID analysis. The criteria for logistic regression and predictive discriminant analysis for a predictor to be entered in the model was a p-level of 0.05 and for a predictor to stay was 0.10.

6.5.3. CHAID: Results and Discussion

The CHAID analysis yielded the decision tree depicted in Figure 6.4.

The CHAID analysis revealed that *Mscore* and *Gender* are the significant predictors of academic success for the BSc students. The CHAID analysis implies a possible interaction between *Mscore* and *Gender*. It also indicated that *Race* and *Year Group* were not significant predictors and did not interact with any of the other significant predictor variables.

Figure 6.4 CHAID Tree Diagram for BSc Group



6.5.4. Stepwise Logistic Regression: Results and Discussion

The results of the logistic regression analysis are given in Table 6.16. This was obtained by entering the variables as described in Section 6.5 as well as an interaction variable between *Mscore* and *Gender* (*Mscore* * *Gender*). *Race* was omitted (see section 4.7.3.1). By entering *Study Orientation* as described in Section 6.1 instead of the variables a different set of variables was selected, namely *Gender*, *Mscore*, *Self-control,* and *Travel*. According to Hosmer and Lemeshow (2000) it is important to find the most parsimonious yet clinical (in this study's case educational) acceptable model in the model building process. The decision

was made to use the model into which the four separate reliable subtests of the SSHA were entered. In Table 6.16 descriptive measures for this model are given. The p-values in Table 6.16 are those obtained for the logistic analysis, in the case of random sampling. The reciprocal of the odds ratio in cases where the odds ratio was less than one were computed, because those variables do have a negative effect on academic success as can also be seen from the sign of the estimate. In the case of *Gender* an adjusted odds ratio was not computed because *Gender* is a nominal variable. Male students were coded as 1 and female students were coded as 0 in this dataset. This information is important when interpreting the odds ratio for gender. From Table 6.16 it is clear that *Gender* is 'negatively' related to academic success which means that according to the logistic procedure if male is coded 1 it is negatively related to academic success.

Table 6.16 Descriptive measures of the stepwise logistic analysis: BSc group

Variable	Estimate	Standard error	Chi-square	P-value	Effect size	Adjusted Odds Ratio	Odds Ratio	1/ Odds Ratio	Correlation coefficient with logit
Intercept	-6.73	1.44	21.99	<.0001					
Delay Avoidance	0.06	0.02	5.86	0.02	0.20	1.72			0.63
Mscore	0.21	0.05	18.66	<.0001	0.35	2.69			0.78
Gender	-1.32	0.44	8.98	0.00	0.25		0.23	4.35	

n = 149 n_1 = 59 n_0 = 90

Three predictor variables were selected with the stepwise selection method, namely *Delay* Avoidance, *Mscore* and *Gender.* The number of parameters p in the model must meet the requirement that $p+1<\min(n_1,n_0)/10=5.9$. For this model it is met because $p+1=4$ (see Sections 3.2.2.9 and 4.7.3.1).

Linearity with the logit

To interpret the odds ratio for continuous variables correctly the requirement is that the relationship between the specific variable and the logit must be linear. The correlation coefficient of each of *Mscore* and *Delay Avoidance* with the logit are practically significant and thus both of them are linearly related to the logit (see Table 6.16).

Three predictor variables were selected with the stepwise logistic regression method, namely *Delay Avoidance*, *Mscore* , and *Gender.* Gender is the predictor variable with the highest

odds ratio, namely 4.35 and is according to Section 3.3.5 highly significant. An odds ratio of 4.35 indicates that for a female student the chance for completing a BSc degree in the prescribed time is 4.35 higher than for a male student, if the other predictors in the model are held constant.

Mscore is the predictor with the second highest odds ratio, namely 2.69 and is substantially significant. An odds ratio of 2.69 indicates that for every increase of 4.81 (one standard deviation) in M*score* the chance for completing a BSc degree in the prescribed time increases by a factor 2.69, if the other predictors in the model are held constant.

Table 6.17 gives the results of the Hosmer and Lemeshow goodness-of-fit test.

Table 6.17 Hosmer and Lemeshow goodness-of-fit test: BSc group

		Status = 1		Status = 0	
Group	Total	Observed	Expected	Observed	Expected
1	15	0	0.64	15	14.36
2	15	1	1.22	14	13.78
3	15	3	1.99	12	13.01
4	15	5	3.3	10	11.7
5	15	4	4.48	11	10.52
6	16	8	6.58	8	9.42
7	15	4	8.35	11	6.65
8	15	11	9.57	4	5.43
9	15	12	11.44	3	3.56
10	13	11	11.44	2	1.56

n = 149 chi-square = 8.98 p-value = 0.34 (in case of random sampling)
w = 0.24 df = 8.

The chi-square value of the Hosmer and Lemeshow goodness-of-fit test was 8.98 and the degrees of freedom 8. The effect size is then w = 0.24 which is a small effect size (defined as of no practical significance) and thus means that the fit of the model is good (see Section 3.3.3). Seven of the cells have expected frequencies of less than 5, which makes the conclusion that the model fits, less valid (Hosmer & Lemeshow, 2000).

Area under ROC curve

The area under the ROC curve was found to be 0.83, which is considered meritorious discrimination (see 3.2.2.9).

Cross validation

The leave-one-out principle was used for cross validation of the logistic procedure. That is, dropping the data of one subject and then re-estimating the parameter estimates to classify that subject. A cut-off point of 0.5 was used for the classification.

Table 6.18 Classification table for stepwise logistic regression: BSc group

		Predicted Status		
		0	**1**	**Total**
Actual Status	**0**	70	20	**90**
		77.78	22.22	**100**
	1	19	40	**59**
		32.20	67.80	**100**
Total		**89**	**60**	**149**
		59.73	**40.27**	**100**

Improvement over chance

The effect size index I for improvement over chance was calculated from Table 6.18. The observed hit rate (H_o) was (91+34)/149 = 0.74 and the expected hit rate (H_e) was 0.52 with n = 149. By using Equation 3.46, I was calculated for the logistic regression model and found to be equal to 0.46, which is a practical significant improvement over chance, suggesting a valid model.

Best subset selection

The three predictors selected as the best subset of three predictors with the logistic regression procedure were *Delay Avoidance*, *Mscore*, and *Gender*. The global score test chi-square value, C, was 46.42 (see Section 3.2.2.8).

6.5.5. Stepwise Predictive Discriminant Analysis: Results and Discussion

To determine whether a linear classification rule or a quadratic rule had to be used, a chi-square test was used to test for equal covariance matrices of the two groups. The chi-square value was 4.43, the degrees of freedom 6 and the p-value = 0.61 (in the case of random sampling). By calculating the effect size w (see Section 3.3.3), the value w = 0.17, which is not practically significant meaning that a pooled covariance matrix and thus a linear classification rule could be used. Table 6.19 gives the coefficients of the linear discriminant functions for the two status groups. The prior probabilities were chosen as 0.39 (39%) for the academic unsuccessful group and 0.61 (61%) for the academic successful group, in accordance with the percentage of failures and successes in the sample.

Table 6.19 Linear Discriminant Function for status: BSc group

Variable	Status = 0	Status = 1
Constant	-20.45	-27.53
Delay Avoidance	0.30	0.36
Mscore	1.28	1.49
Gender	4.59	3.19
Priors	**0.61**	**0.39**

n = 149 n_1 = 59 n_0 = 90

Three predictors were selected with the stepwise selection procedure, namely *Delay Avoidance*, *Mscore*, and *Gender*. According to Section 3.2.3.2 for a selection of five predictors $n_j > 3p$, as $n_j = 59$ and $3p = 9$ and the requirement is thus met.

Cross validation

The leave-one-out principle was used for cross validation of the discriminant analysis procedure. An observation was classified by calculating its value for both linear discriminant functions (based on the data without that observation) and then classified into the group for which the value of the function was the highest. Table 6.20 gives the classification according to the predictive discriminant analysis.

Table 6.20 Linear Discriminant Function for status: BSc group

	Predicted Status		
Actual Status	**0**	**1**	**Total**
0	69	21	**90**
	76.67	23.33	**100**
1	21	38	**59**
	35.59	64.41	**100**
Total	**90**	**59**	**149**
	60.40	**39.60**	**100**

Improvement over chance

For the discriminant analysis model the observed hit rate (H_o) was (69+38)/149 = 0.72 and the expected hit rate (H_e) was 0.52 with n = 149 (from Table 6.20). When *I* was calculated for this model it was found to be equal to 0.41, which is also a practical significant improvement over chance (see Section 3.3.5), indicating a valid model.

6.5.6. Evaluation

The same three predictors that were selected by the stepwise logistic regression, *Delay* Avoidance*, Mscore, and Gender* were also selected by stepwise discriminant analysis for the BSc group. The three predictors that were selected as the best subset selection in logistic regression were also similar to that of the stepwise logistic and discriminant analysis' stepwise selection procedures. However, only *Mscore* and *Gender* were selected by the CHAID procedure. *Mscore* and Gender were thus selected by all four model fitting procedures.

The sensitivity of the model fitted by logistic regression is 67.80% (40/59) and the specificity is 77.78% (70/90), as can be seen from Table 6.18. The sensitivity of the model fitted by predictive discriminant analysis is 64.41% (38/59) and the specificity is 76.67% (69/90), as can be seen in Table 6.20. When comparing the students who were misclassified by logistic regression, with those who were misclassified by predictive discriminant analysis, there were three students who were misclassified by discriminant analysis, but correctly classified by logistic regression. Two of these students were wrongly classified by discriminant analysis

as failures, while they were actually successes. This result is confirmed by the fact that the specificity of the discriminant analysis is approximately three percent lower than the specificity of logistic regression. One was wrongly classified by discriminant analysis as a failure, but was a success which is the reason for the specificity to be one percent lower.

Chapter 7

7. Conclusions, limitations and recommendations

In this Chapter the four research questions will be discussed by bringing together the results reported in Chapters 5 and 6.

7.1. Psychometric tests

The first and second research questions are the following: Are the SAT 78, GSAT, SSHA, PHSF, and 19 FII

1. reliable instruments for the study sample and, if so, how does the reliability of these instruments compare with the reliability at the time of their standardisation?
2. construct valid instruments for the study sample and, if so, how does the construct validity of these instruments compare with the construct validity at the time of their standardisation?

7.1.1. SAT 78

The number of respondents to determine reliability and construct validity for the Study Sample (2003-2007) was 2 084 while it was 1 453 in 1978.

7.1.1.1. Reliability

Satisfactory to high Cronbach alpha coefficients were obtained for all 10 subtests for the Study Sample (2003-2007), and the values compare favourably with the K-R 8 values obtained in 1978 for the 10 different subtests. This means that the SAT 78 is still a reliable test for the Study Sample (2003-2007). These results are reported in Table 5.2.

7.1.1.2. Construct validity

From the results reported by Fouche & Verwey (1978) and shown in Table 5.3 four factors were found in the 1978 exploratory factor analysis. These four factors formed the four constructs that were given names such as *Verbal Ability* by the constructors of the SAT 78.

In contrast, only three factors (which may form constructs) were found in this study (see Table 5.4). The constructs included different combinations of subtests with different factor loading and were hence incomparable to the original constructs.

Calculations and *Comparison* load highly on different constructs in this study, although, according to Fouche & Verwey (1978), they should form the construct *Numerical Ability*. This means that the *Numerical Ability* construct was not evident in the Study Sample (2003-2007).

As a result of the high loadings of the subtests *Disguised Words*, *Verbal Comprehension*, and *Comparison*, together with *Memory (Paragraph)* and *Memory (Symbols)* on construct 2 of this Study Sample (2003-2007), construct 4 (*Memory*) of the original SAT 78 is also not present on its own in the Study Sample (2003-2007). Although *Memory (Paragraph)* and *Memory (Symbols)* load on the same construct in the Study Sample (2003-2007), it is not possible to combine these two subtests in the Study Sample (2003-2007) to form a construct *Memory* on its own similar to that of the SAT 78.

There is a high loading of the subtest *Figure Series* on its own as a one variable construct in the Study Sample (2003-2007) while it is part of construct 3 (*Visual-Spatial Reasoning*) of the SAT 78. The conclusion is made that the SAT 78 is not a construct valid instrument for the Study Sample (2003-2007).

These results for the exploratory factor analysis, which were very different to those of the Fouche and Verwey (1978) study, were obtained even though all the conditions of Kaiser's MSA, and sample size were fulfilled. The restrictions on skewness and kurtosis of the data were also more than satisfactorily met.

7.1.2. GSAT

The number of respondents used to determine reliability and construct validity for the Study Sample (2003-2007) was 591 while it was 138 in 1978. The sample of 138 used in 1991 was respondents in grade 12 as reliability coefficients were reported separately per grade in the manual (Claassen *et al.*, 1991). The number of respondents to determine construct validity for the Study Sample (2003-2007) was also 591 while it was 786 in 1991 for respondents from grade 8 to grade 12.

7.1.2.1. Reliability

High Cronbach alpha coefficients were obtained in this study for both the subtests and the constructs. The reliability coefficients compare very well with the K-R 8 obtained in 1991 (see Table 5.6). Thus, the GSAT remains a highly reliable test for the Study Sample (2003-2007).

7.1.2.2. Construct Validity

A verbal and non-verbal construct could not be separated for either the Study Sample (2003-2007) or at the time of standardisation, although throughout the manual of the GSAT the *Verbal* and *Non-verbal* constructs are separated as if loading separately (Claassen *et al.*, 1991). One construct, presumably *Total,* was retained for the GSAT on the Study Sample (2003-2007) and one was retained at the time of standardisation as reported in Tables 5.8 and 5.9. This result indicates that the GSAT remains a construct valid instrument for the Study Sample (2003-2007) in the sense that *Total* is the one and only construct.

7.1.3. SSHA

The number of respondents to determine reliability and construct validity for the Study Sample (2003-2007) was 2 084 while it was 1 453 in 1974.

7.1.3.1. Reliability

High Cronbach alpha coefficients were obtained for the Study Sample (2003-2007) and the values compare very well with the split-half reliability coefficients obtained in 1974 for the different subtests of the SSHA (see Table 5.12). This means that the SSHA is a highly reliable test for the Study Sample (2003-2007).

7.1.3.2. Construct validity

Construct validity was not determined for the SSHA at the time of standardisation. By using the four subtests *Delay Avoidance*, *Work Methods*, *Teacher Approval*, and *Education Acceptance* as the variables in an exploratory factor analysis in this study, one construct was retained for the SSHA on the Study Sample (2003-2007), presumably *Study Orientation*. It was not possible to distinguish between the different facets of study habits, namely *Study Habits* and *Study* Attitude, because these two constructs were not extracted by the factor analysis for the Study Sample (2003-2007). The results, however, seem favourable to regard the SSHA as a construct valid instrument for the Study Sample (2003-2007) in the sense that *Study Orientation* is the one and only construct. This is reported in Table 5.13.

7.1.4. PHSF

The number of respondents to determine reliability and construct validity for the Study Sample (2003-2007) was 2 585 while it was 1 788 in 1983.

7.1.4.1. Reliability

Satisfactory to high Cronbach alpha coefficients were obtained in this study and the values compare very well with the split-half values obtained in 1983 for the different subtests. This means that the PHSF remains as a reliable test for the Study Sample (2003-2007). This is reported in Table 5.15.

7.1.4.2. Construct Validity

The constructs of the PHSF for the study sample differ from those found at the time of standardisation. Three constructs were retained for the Study Sample (2003-2007). The original test's construct *Home Relations* was also found in this study. However, the other two constructs found were indefinable constructs with hardly any similarity to any of the seven other constructs retained at the time of standardisation. The conclusion is made that the PHSF is not a construct valid instrument for the Study Sample (2003-2007) (see Tables 5.16 and 5.17). These results for the exploratory factor analysis, which were very different to those of the Fouche & Grobbelaar (1983) study, were obtained even though all the conditions of Kaiser's MSA and sample size had been fulfilled. The restrictions on both skewness and kurtosis of the data were more than satisfactorily met.

7.1.5. 19 FII

The number of respondents used to determine reliability and construct validity for the Study Sample (2003-2007) was 2 597 while it was 903 in 1977.

7.1.5.1. Reliability

High Cronbach alpha coefficients were obtained in this study and the values compare very well with the split-half values obtained in 1977 for the different subtests. The 19 subtests of the 19 FII were found reliable for the Study Sample (2003-2007) and compare well to the reliability at the time of standardisation. The two additional subtests namely *Work-Hobby* and *Active-Passive* had Cronbach alpha coefficients of less than 0.70 indicating that these two subtests are not reliable subtests for the study sample. This is reported in Table 5.19.

7.1.5.2. Construct Validity

The constructs of the 19 FII for the Study Sample (2003-2007) differ very much from those found at the time of standardisation. Six constructs were retained in this study, incomparable with the groupings of the study of 1997. The conclusion is made that the 19 FII is not a

construct valid instrument for the Study Sample (2003-2007) (see Tables 5.20 and 5.21), although all the restrictions of Kaiser's MSA, sample size, skewness, and kurtosis were met.

7.1.6. Conclusion

By examining the reliability of the five psychometric tests for the Study Sample (2003-2007) a positive conclusion can be made that except for the two subtests *Work-Hobby* and *Active-Passive* of the 19 FII the tests have remained reliable instruments.

The picture of the tests' construct validity is not that satisfactory. The constructs of the SAT 78, PHSF, and 19 FII in this study differ materially from those at the time of standardisation. This conclusion is made even though all the restrictions were met when the factor analyses were done on the study samples. The constructs of the SAT 78, PHSF and 19 FII could thus not be used as predictor variables for the model fitting techniques. Neither could the *Verbal* and *Non-verbal* constructs of the GSAT nor the constructs *Study Habits* and *Study Attitude* of the SSHA be used. The subtests *Total* of the GSAT and *Study Orientation* of the SSHA could be used as predictor variables. *IQ* may also be used because it is calculated from reliable subtests of the SAT 78.

7.2. Models

The third and fourth research questions in this study are the following:

3. Which of the available reliable and construct valid predictors are the best at predicting academic success for BCom, BPharm, BA, and BSc students, respectively?
4. Are there models which can adequately predict academic success for each of the BCom, BPharm, BA, and BSc degrees, respectively?

7.3. Interpretation of the Results of the fitted Logistic Regression Models

In this study four stepwise logistic regression models, one for each of the BCom, BPharm, BA, and BSc groups had been fitted. These four models had been validated by using a best subset selection in logistic regression, CHAID and discriminant analysis (see Chapter 6). Interpretation of the results of the fitted models in this study is being done by using the odds ratios of the predictors (which were obtained from the estimates of the coefficients) of the models built with the stepwise logistic procedure. It is important to note that the interpretation of the odds ratios of a predictor is in the context of a specific model, that is in combination with other predictors and not in isolation.

7.3.1. Bachelor of Commerce

The significant predictors selected by the **CHAID** decision tree are *Mscore, Health, Business, Practical-female*, *Law, and Public Speaking*. *Mscore* is the most significant predictor according to the CHAID procedure.

The **stepwise logistic regression analysis**, the **best subset of seven predictors of the logistic regression** procedure as well as **stepwise discriminant analysis** selected *Self-control, Social Work, Public Speaking, Law, Disguised Words, Figure Series,* and *Mscore*. *Self-control, Social Work, Figure Series,* and *Mscore* are positively related to academic success because their odds ratios are above 1. Each of *Public Speaking, Law,* and *Disguised Words* are negatively related to academic success because their odds ratios are less than 1 (see Table 6.1). This interpretation is made in the context of the whole model.

Disguised Words and *Figure Series* are subtests of the SAT 78. *Health* and *Self-control* are subtests of the PHSF while *Business, Practical-female*, *Law, and Public Speaking* are subtests of the 19 FII.

None of the subtests of the SSHA were selected by any of the procedures used in this study to predict academic success for the BCom group. *IQ* was also not selected as a predictor of academic success.

Mscore, Public Speaking, and *Law* were selected by all four procedures (that is stepwise logistic regression, best subset in logistic regression, discriminant analysis and CHAID). However, the only predictor according to the stepwise logistic regression procedure which is of practical importance is *Mscore.*

Thus, to answer the question of which are the best predictors for academic success for the BCom group, *Self-control, Social Work, Public Speaking, Law, Disguised Words, Figure Series,* and *Mscore* came out in combination as the best predictors.

As discussed in Section 3.2.2.9 **classification tables** are a very appealing way to summarise the results of a fitted logistic regression model, but accurate or inaccurate classification does not address the criteria for goodness-of-fit (Hosmer & Lemeshow, 2000). Classification depends on the choice of a cutpoint for the probability of a success, which was in this study's case 0.5, because of the relationship between logistic regression and discriminant analysis (see Section 3.2.2.9). Classification is important when, as in this study, it is essential to

classify students in a success or failure group. It should, however, just compliment methods of assessment of fit (Hosmer & Lemeshow, 2000).

From the information supplied by the classification table (Table 6.3), when the leave-one-out method was used, the sensitivity of the model fitted by logistic regression was computed and found to be 75.73% and for the model fitted by discriminant analysis it was found to be 75.24%. The specificity for the logistic model is 58.81% and for the discriminant analysis it is 53.81%. In the environment of tertiary education, students who do not pass are financially very expensive both to themselves, the university, and the state. Students who do not pass are financially more problematic for the university than it is beneficial to the university if the students pass their courses. Thus, the misclassification cost is higher when a student is wrongly classified as a success, when in fact he or she is a failure. It is thus a disadvantage of this model that the specificity is lower than the sensitivity (Huberty & Olejnik, 2006). The improvement over chance index, *I,* for the stepwise logistic analysis was equal to 0.34 which is practically significant and for stepwise discriminant analysis *I* = 0.30, which is a medium to large effect of improvement over chance. Both procedures supported each other by selecting the same predictor variables.

According to the Hosmer and Lemeshow **goodness-of-fit-test** the fit of the model was valid, as reported in Table 6.2, $w = 0.13$. This means that the probabilities are reflecting the true outcome in the data, that is the model is well calibrated. The area under the **ROC** curve was 0.75 and indicates that the model has acceptable discrimination (see Section 3.2.2.9). According to Hosmer & Lemeshow (2000) these two criteria namely calibration and discrimination are the most important when accessing model performance. It is, however, possible that a poorly fitting model may still have good discrimination (Hosmer & Lemeshow, 2000).

Thus, a model was found which adequately fits the data and has acceptable discrimination. This model can also adequately predict academic success because the improvement over chance was practically significant for the logistic regression procedure Limitations are discussed in Section 7.4.

The suggested stepwise logistic regression model for the BCom students is:

$$\mathrm{P}(\text{success}) = \frac{e^{-3.44 + 0.04\,Self\text{-}control + 0.03\,Social\ Work - 0.03\,Public\ Speaking - 0.02\,Law - 0.04\,Disguised\ Words + 0.02\,Figure\ Series + 0.16\,Mscore}}{1 + e^{-3.44 + 0.04\,Self\text{-}control + 0.03\,Social\ Work - 0.03\,Public\ Speaking - 0.02\,Law - 0.04\,Disguised\ Words + 0.02\,Figure\ Series + 0.16\,Mscore}}.$$

7.3.2. Bachelor of Pharmacy

The significant predictors selected by the **CHAID** decision tree were *Mscore, Chemistry,* and *Practical-male. Mscore* is the most significant predictor according to the CHAID procedure.

The **stepwise logistic regression** analysis, the **best subset of five predictors of logistic regression** procedure as well as **stepwise discriminant analysis** selected *Family Influences, Science, Public Speaking, Law,* and *Mscore*. *Family Influences, Science, Public Speaking, and Mscore* are positively related to academic success, because their odds ratios are above one. *Law* is negatively related to academic success because its odds ratio is less than one. This interpretation is made in the context of the whole model.

Family Influences is a subtest of the PHSF, while *Science, Public Speaking, Law,* and *Practical-male* are subtests of the 19 FII. The *Chemistry* test, which was part of three tests for the selection of the BPharm students, was only selected by the CHAID procedure as a significant predictor.

None of the subtests of the GSAT and SSHA were selected by any of the procedures used in this study to predict academic success for the BPharm group. The *Mathematics* and *Physics* tests administered for selecting the students for the BPharm degree were also not selected by any of the procedures. The subtest *Total*, which is the subtest of the GSAT that gives an indication of a testee's intelligence, was also not selected as a predictor of academic success for the BPharm group.

Mscore and *Law* were selected by all four procedures. The predictor with the highest odds ratio according to the stepwise logistic regression procedure was *Mscore*, although an odds ratio of 1.73 is not practically significant.

Thus, to answer the question of which are the best predictors at predicting academic success for the BPharm group, *Family Influences, Science, Public Speaking, Law,* and *Mscore* came out in combination as the most important predictors, although none of their odds ratios were practically significant.

The logistic regression and discriminant analysis procedures supported each other by selecting the same predictor variables. From the information supplied by the **classification table**, when the leave-one-out -method was used, the sensitivity of the model fitted by logistic regression was computed and found as 96.46% and for the model fitted by discriminant

analysis it is 95.54%. The specificity for the logistic model is 10.53% and for the discriminant analysis it is 8.62%. Classification is sensitive to the relative sizes of the two groups and in the case of this group the failure group (22.35%) is small relative to the success group (77.65%). Furthermore, classification always favours the larger group, which is thus the reason for the high sensitivity and low specificity of this group (Hosmer and Lemeshow 2000). The low specificity obtained by both the stepwise logistic regression and discriminant analysis models is, however, highly problematic. This implies that the model does not have the ability to predict an academic failure well. Although the sensitivity of both models is excellent, that is that the models can predict success well, there is a tendency for both of the models to predict a student as a success while he or she is actually a failure. It would be more beneficial for the university to know if a student is at risk to be a failure, than it is to predict if a student is a potential success as a result of the fact that the misclassification cost is higher when the student is a failure than when he or she is a success (see Section 7.3.1). The improvement over chance index, *I,* for the stepwise logistic analysis was equal to 0.34 and for stepwise discriminant analysis *I* = 0.31. Both these indices indicate that these models predict academic success with a medium to large practical effect, better than chance.

As stated in Section 7.3.1 classification tables do not address the criteria for goodness-of-fit. That means that a model can be poorly calibrated but still predict group membership well (Hosmer & Lemeshow, 2000).

According to the Hosmer and Lemeshow **goodness-of-fit** test the fit of the model, although *w* = 0.17, was possibly not valid because there were five cells with expected frequencies of less than 5, as reported in Table 6.7. The area under the **ROC** curve was 0.72 which is an indication of acceptable discrimination of the model.

Thus, a model could be found with acceptable discrimination, but with its goodness-of-fit in question. This model could adequately predict academic success for the BPharm group, as a result of the fact that the improvement over chance index, *I,* is nearly practically significant.

The respondents of the study sample used for the model fitting procedures of the **BPharm** group were a selected high achieving academic group. The School of Pharmacy is considered as one of the best pharmacy schools in South Africa, producing the most undergraduate and graduate students as well as research publications in its discipline in the country. These students are the group of students at this university with the highest average *Mscore* of all courses at this university (see Sections 6.2.1, 6.3.1, 6.4.1, 6.5.1). In 2003 and 2004 there were a large number of applicants for the BPharm degree and their *Mscore* were

used as the most important factor in selection. Furthermore, they had to pass both mathematics and physical science with at least an E symbol on higher grade or at least a D symbol on standard grade. The fact that this group of pharmacy students was already a highly selected academic group must be kept in mind when implementing this model. Limitations are discussed in Section 7.4.

The suggested stepwise logistic regression model is:

$$P(\text{success}) = \frac{e^{-6.41 + 0.07\,Family\ Influences + 0.03\,Science + 0.03\,Public\ Speaking - 0.04\,Law + 0.16\,Mscore}}{1 + e^{-6.41 + 0.07\,Family\ Influences + 0.03\,Science + 0.03\,Public\ Speaking - 0.04\,Law + 0.16\,Mscore}}.$$

7.3.3. Bachelor of Arts

The significant predictors selected by the **CHAID** decision tree were *Practical-Male*, *Spatial 2D,* and *Numerical* were the significant predictors of academic success for the BA students.

The **stepwise logistic regression** analysis, the **best subset of five predictors of the logistic regression** procedure as well as **stepwise discriminant analysis** selected *Self-control, Nature, Comparison, Spatial 2D,* and *Mscore. Self-control, Comparison, Spatial 2D,* and *Mscore are* positively related to academic success, because their odds ratios are above one. *Nature* is negatively related to academic success because its odds ratio is less than one, as reported in Table 6.11. This interpretation is made in the context of the whole model.

Self-control is a subtest of the PHSF, while *Nature* is a subtest of the 19 FII. *Comparison* and *Spatial 2D* are subtests of the SA 78.

None of the subtests of the SSHA was selected by any of the procedures used in this study to predict academic success for the BA group. *IQ* was also not selected as a predictor of academic success for the BA group.

Spatial 2D was selected by all four procedures. The predictor with the highest odds ratio according to the stepwise logistic regression procedure was *Self-control,* namely 1.82 although an odds ratio of this magnitude is not practically significant.

Thus, to answer the question of which are the best predictors at predicting academic success for the BA group *Self-control, Nature, Comparison, Spatial 2D,* and *Mscore* came out as the

most important predictors in combination, but none of their odds ratios are practically significant.

From the information supplied by the **classification table**, when the leave-on-out –method was used, the sensitivity of the model fitted by logistic regression was computed and found as 53.12% and for the model fitted by discriminant analysis it is 51.56. The specificity for the logistic model is 84.26% and for the discriminant analysis it is 82.73. The fact that the misclassification cost is higher when a student is classified wrongly as a success, when in fact he or she is a failure, is thus a benefit of this model, because the specificity is higher than the sensitivity (see Section 7.3.1). The improvement over chance index, *I,* for the stepwise logistic analysis was equal to 0.43 and for stepwise discriminant analysis *I* = 0.38 and were both large effects and therefore practically significant. Both procedures supported each other by selecting the same predictor variables.

According to the Hosmer and Lemeshow **goodness-of-fit** test although $w = 0.19$ the fit of the model may not be valid because there were five cells with expected frequencies less than five, as shown in Table 6.12. The area under the **ROC** curve was 0.75 which is considered acceptable discrimination.

The conclusion is made that a model could be found with acceptable discrimination, although the conditions for the goodness-of-fit test may not have been met. Furthermore, this model could adequately predict academic success for the BA group with a practically significant improvement over chance. Limitations are discussed in Section 7.4.

The suggested stepwise logistic regression model is:

$$\mathrm{P}(\text{success}) = \frac{e^{-9.70 + 0.12\,Self\text{-}control - 0.05\,Nature + 0.14\,Comparison + 0.08\,Spatial\,2D + 0.09\,Mscore}}{1 + e^{-9.70 + 0.12\,Self\text{-}control - 0.05\,Nature + 0.14\,Comparison + 0.08\,Spatial\,2D + 0.09\,Mscore}}.$$

7.3.4. Bachelor of Science

The significant predictors selected by the **CHAID** decision tree were *Mscore,* and *Gender.*

The **stepwise logistic regression** analysis, the **best subset of three predictors of the logistic regression** procedure as well as **stepwise discriminant analysis** selected *Delay Avoidance*, *Mscore* , and *Gender.*

Mscore and *Delay Avoidance* are positively related to academic success because their odds ratios are above one. *Gender* is negatively related to academic success because the odds ratio is less than one which implies that it is 4.35 times more likely to be a failure for a male student than for a female student, as reported in Table 6.16. This interpretation is made in the context of the whole model.

Delay Avoidance is a subtest of the SSHA.

None of the subtests of the SAT 78, PHSF or 19 FII was selected by any of the modelling procedures used in this study to predict academic success for the BSc group. *IQ* was also not selected as a predictor of academic success for the BSc group.

Mscore and Gender were selected by all four procedures. The predictor with the highest odds ratio according to the stepwise logistic regression procedure was *Gender* with an odds ratio of 4.35 which is highly significant and *Mscore* with an adjusted odds ratio of 2.69 which is substantially significant (see Table 6.16).

CHAID implies an interaction between *Mscore* and *Gender,* but stepwise logistic regression did not select the interaction as a possible predictor, when entering it as a predictor in the stepwise logistic regression procedure.

The best predictors at predicting academic success for the BSc group is *Gender, Mscore,* and *Delay Avoidance. Gender* and *Mscore* have odds ratios which is practically significant. The odds ratio of *Delay Avoidance* was not practically significant.

The sensitivity of the model fitted by logistic regression is 67.80% and for the model fitted by discriminant analysis it is 64.41%. The specificity for the logistic model is 77.78% and for the discriminant analysis it is 76.78%. The fact that the misclassification cost is higher when a student is classified wrongly as a success, when in fact he or she is a failure, is thus a benefit of this model, because the specificity is higher than the sensitivity (see Section 7.3.1). These values were computed from the **classification tables**. The improvement over chance index, *I,* for the stepwise logistic analysis was equal to 0.46 and for stepwise discriminant analysis *I* = 0.41 and both are practically significant. Both procedures supported each other by selecting the same predictor variables.

According to the Hosmer and Lemeshow **goodness-of-fit** test the fit of the model may not be valid because there were seven cells with expected frequencies less than five, although w = 0.24. This is reported in Table 6.17. The area under the **ROC** curve is 0.83 which is considered meritorious discrimination. Limitations are discussed in Section 7.4.

Thus, a model could be found with meritorious discrimination, but with its goodness-of-fit in question. This model could adequately predict academic success for the BSc group, as a result of the fact that the improvement over chance index I is practically significant. The fact that this group of BSc students was already a selected academic group because they had passed mathematics in matric on higher grade, must be also kept in mind when implementing this model.

The suggested stepwise logistic regression model is:

$$P(\text{success}) = \frac{e^{-6.73 + 0.06\,Delay\ Avoidance + 0.21 Mscore - 1.32 Gender}}{1 + e^{-6.73 + 0.06\,Delay\ Avoidance + 0.21 Mscore - 1.32 Gender}}.$$

7.3.5. Conclusions

The first conclusion that can be made from this study is that by using biographical, academic history and psychometric test data as predictors, acceptable models could be found to predict academic success for all four groups of students in this study.

Different predictors as well as different models were found by the four procedures used in this study for the four different degree types. This means that for different courses different predictors were selected to predict academic success. An interesting fact is that no aptitude subtests were selected for either the BSc or BPharm group. That means that none of the subtests of the SAT 78 were selected for the BSc group and none of the subtests of the GSAT were selected for the BPharm group. In contrast, three of the subtests of the 19 FII (*Science, Public Speaking,* and *Law*) and one of the subtests of the PHSF (*Family Influences*) were selected as predictors for the BPharm group. Furthermore, *Delay Avoidance*, a subtest of the SSHA, was selected as a predictor for academic success for the BSc group. These two groups were already selected academic groups (see Sections 7.3.2 and 7.3.4) which seems that interests, influences, and study methods begin to count for academic achievement in the presence of academic expertise.

On the other hand for the BCom and BA groups, whose admission requirements to this university are not as strict as those of the BPharm and BSc groups, subtests of the SAT 78 like *Disguised Words* and *Figure Series* for the BCom group and *Comparison* and *Spatial 2D* for the BA group were selected. It thus seems that in the absence of strict academic admission criteria, aptitudes like these are found to be important predictors for academic success.

Mscore was the only predictor which was selected by all the procedures for all the groups with the exception of the CHAID procedure for the BA group. *Mscore* in combination with other predictors can thus be regarded as the single most important predictor of academic success found by this study. *Public Speaking* and *Law* are both selected by the stepwise logistic procedure, the best subset logistic regression procedure and the discriminant analysis to predict academic success for both the BCom and BPharm groups. The fact that *Mscore* was selected for the BPharm group as a predictor is fascinating, because as mentioned in Section 7.3.2, these students' *Mscore* were already taken into account when they were selected for this course. This means that the variation in *Mscore* among these students was already lessened. This fact is surely an indication of how important matric results as a predictor of academic success seems to be on this campus.

IQ which was computed from subtests of the SAT 78 for the BCom, BA, and BSc groups was not selected as a predictor by any of the procedures for any of these groups. The *Verbal* subtest of the GSAT, which is a measure of *IQ*, an indication of general intelligence, was not selected for the BPharm group as a predictor of academic success.

Gender was selected for the BSc group only and is the only predictor with a highly significant odds ratio in all the models. The odds ratio for gender for the BSc group indicates that it is 4.35 times more likely for a female to obtain a BSc degree in the minimum time than for a male if the other predictors in the model are kept constant. *Mscore* had a practical significant adjusted odds ratio for the BCom and BSc groups. The odds ratio for *MScore* for the BCom indicates that for every increase of 4.85 (one standard deviation, see Section 4.7.3.1) in *Mscore* the chance for completing a BCom degree in the prescribed time increases by a factor 2.2, if the other predictors in the model are kept constant. The odds ratio for Mscore for the BSc group indicates that for every increase of 4.81 (one standard deviation) in *Mscore* the chance for completing a BSc degree in the prescribed time increases by a factor 2.69, if the other predictors in the model are kept constant.

According to Hosmer and Lemeshow (2000) no comparisons about sensitivity or specificity could be made across models.

The model with the best discrimination is the one for the BSc group, with a discrimination value of 0.83, which is considered meritorious. In the case of this group there were 59 successes and 90 failures, thus 149 respondents in total. If each respondent with y = 1 (a success) is paired with each respondent with y = 0 (a failure), then 59 x 90 = 5310 pairs are created. By counting the number of times that a subject y = 1 had a higher probability than one with y = 0 and dividing it by the total number of pairs, the ratio 4407/5310 = 0.83 is found, which is then similar to the area under the ROC curve. That means that this model classified a success correctly as a success 4 407 times out of 5 310 times (see Section 3.2.2.9).

When looking at the different models, and taking calibration (i.e. the probabilities reflecting the true outcome experience in the data) and differentiation of the model into account, the model for the BCom group was the best. The area under the ROC curve of this model was 0.75, which is acceptable discrimination. According to the Hosmer and Lemeshow goodness-of-fit test, this model fits the data well, because w = 0.13 (p = 0.59 in case of random sampling). Furthermore the restriction about a minimum of 5 expected frequencies in each cell in the 2 x 10 frequency table is met, which is not the case for the other three models found in this study.

7.3.6. Limitations

Although it is beneficial to have the opportunity to use available data for research as mentioned in Chapter 1, it has its limitations.

In the case of this study the planning had to be done around the available data. It may have been preferable to have more biographical as well as psychometric data for every student, because academic success can be related to several facets of a human being. According to Tinto (1975) and Pascarella and Terenzini (1983) academic persistence is strongly related to a student's level of academic and social integration with an institution, commitment to earning a degree, and commitment to an institution. Liu (2000) stated that commonality between integration and satisfaction is crucial to the success of academic performance and persistence and that students' satisfaction is highly related to student retention. This information was not available. As no information on, for example, the socio-economic status or the academic background of the parents was available; this shortage of personal motivational and familial data is a possible limitation of using these available data.

The Potchefstroom Campus of the North West University has predominantly white students. This study could not give any answers regarding the reliability and construct validity of these psychometric tests for black, coloured or Indian students.

All the study samples on which the models had been fitted consisted of predominantly white students. If these models would be used in future and the race of the students is more diverse than for the study samples, it is possible that the prediction of the models could be different.

The respondents of the **BCom**, **BA**, and **BSc** groups in the study samples used for the model fitting procedures were volunteers. According to Rosenthal and Rosnow (1975) volunteers tend to have certain characteristics, such as being female, firstborn, sociable, extroverted, etc. This fact could thus be seen as a limitation of this study, because if having more diverse groups (that is in case of compulsory testing), the models found for the BCom, BA and BSc groups may predict academic success differently.

The **BPharm** group was a highly selected academic group. To find predictors to predict academic success in these circumstances where operating in a restricted range can be limiting. To differentiate between predictors in predicting academic success, in the case where a selection of the best possible criteria had already been made, may be problematic because much of the variation in the data had already been removed. In the sense of finding a model to predict academic success for a group like this, it is a limitation. The large discrepancy between the size of the failure group (22.35%) and that of the success group (about 77.65%) can be seen as a limitation for the model fitting of this group. According to Hosmer and Lemeshow (2000) classification always favours the larger group, which is thus the reason for the low specificity of this group, and thus this remains a limitation. This was discussed in Section 7.3.2.

The sample sizes for the BCom, BPharm, BA, and BSc groups were too small for the **CHAID** procedure to be used to the optimum. As a result of this fact the depth of the trees built by CHAID was restricted to 3. The implication was that in this study CHAID did not have a fair chance to be compared with either stepwise logistic regression or stepwise discriminant analysis as a model building technique and this can thus be seen as a limitation (see Section 3.2.1.4).

7.4. Recommendations

Re-evaluation and standardisation of the constructs of the SAT, GSAT, SSHA, PHSF, and 19FII can be considered and standardised for the application on tertiary students of South Africa.

The models found in this study were validated by the leave-one-out method in logistic regression analysis and discriminant analysis. Although this method is considered as an external validation method (Huberty & Olejnik, 2006), it would be useful to see how these models would be if tested on future cohorts. Once available, the graduation data of future cohorts can be used for the BCom, BA, BSc groups (3 year degrees) and BPharm group (4 year degree).

The battery of tests used by this university for guidance and selecting purposes may well need to be revised. Only a few subtests from the psychometric tests were selected by logistic regression and discriminant analysis as predictors of academic success.

In light of the fact that neither *IQ* (SAT) nor *Verbal* (GSAT) were selected by any of the models to predict academic success for any of these groups, other psychometric tests with constructs which measures general intelligence may well be considered, for example the Differential Aptitude Test (DAT) (Foxcroft *et al.* 2004; Owen & Vosloo, 1999).

If a potential student could fill in a questionnaire with personal information about socio-economic status, parents' academic background, as well as completing an appropriate measure on motivation on the same day when the psychometric tests are done, it may improve the prediction of academic success. Some of this information has historically been gathered by the university's Public Relations Department, but it could not be merged with other information of a specific student. Ludeman (2006) reports that one of the goals of the Kellogg Foundation is to gather data on the profile of a student at institution level that can contribute to a national databank of information about academic success or retention at tertiary education level.

By accumulating more data year by year neural networks may be employed to fit models to predict academic success, because neural networks require large data sets. CHAID decision trees could then also be used to help fit models and therefore not only be used for exploratory purposes.

The findings of this study can be used for planning a whole new scenario concerning psychometric testing at the North-West University and nationally for predicting academic success or for selection purposes.

8. Bibliography

Aiken, L.R. & Groth-Marnat, G. (2006). Psychological testing and assessment, 12th ed. Boston: Pearson Education Group, Inc. 535 p.

Anastasi, A. & Urbina, S. (1997). *Psychological testing*, 7th ed. New York: Prentice Hall, Inc. 721 p.

Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York, John Wiley & Sons. 292 p.

Botha, A. (1989). Test anxiety subscale of the achievement motivation test: factor structure. *South African Journal of Education,* vol. 9, no. 1, pp. 22-25.

Budd, J.M. (1988). A bibliometric analysis of higher education literature. *Research in Higher Education,* vol. 28, no. 2, pp. 180-190.

Chen, C. (2005). Analyzing Student Learning Outcomes: Usefulness of Logistic and Cox Regression Models. AIR Professional File, pp. 1-19.

Claassen, N.C.W., de Beer, M., Hugo, H.L.E. & Meyer, H.M. (1991). *Handleiding vir Algemene Skolastiese Aanlegtoets (ASAT) Senior Reeks*. Pretoria, HSRC, 128 p.

Clark, L.A. & Watson, D. (1995). Constructing validity: basic issues in objective scale development, *Psychological Assessment,* vol. 7, no. 3, pp. 309-319.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences,* 2nd ed. Hillsdale, N.J., Erlbaum, 567 p.

de Bruin, G.P. (1997). Spearman se G en die faktorstrukture van die senior aanlegtoetse en die algemene skolastiese aanlegtoets. *Journal of Industrial Psychology*, vol. 23, no. 2, pp. 14-18.

Department of Education, South Africa. (2001). National Plan for Higher Education (NPHE).

de Vetta, H.M. (1987). Differences in performance of black and white students on the Brown-Holtzman survey of study habits and attitudes (SSHA). *South African Journal of Education*, vol. 7, no. 2, pp. 145-150.

du Toit, L.B.H. (1974). *Handleiding vir die opname van studiegewoontes en -houdings (OSGH), Vorm H*. Pretoria, Institute for Psychological and Edumetric Research, HSRC.

Eason, S.J. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In Thompson B. (ed.), *Advances in educational research: Substantive findings, methodological developments,* vol. 1, JAI Press, Greenwich, CT. pp. 83-98.

Eiselen, R. & Geyser, H. (2003). Factors distinguishing between achievers and at risk students: a qualitative and quantitative synthesis. *South African Journal of Higher Education*, vol. 17, no. 2, pp. 118-130.

Ellis, S.M. (2002). The distribution of the residuals of financial risk models. PhD Thesis, Potchefstroom University for Christian Higher Education. 223 p.

Ellis, S.M. & Steyn, H.S. (2003). Practical significance (effect sizes) versus or in combination with statistical significance (p-values). *Management Dynamics*, vol. 12, no. 4, pp. 51-53.

Engelbrecht, M. (1999). *Leerpotensiaal as voorspeller van akademiese sukses van universiteitstudente.* PhD Thesis, Potchefstroom University for Christian Higher Education. 327 p.

Fan, X. & Wang, L. (1999). Comparing linear discriminant function with logistic regression for the two-group classification problem. *Journal of Experimental Education*, vol. 67, no. 3, pp. 265-286.

Field, A. (2005). *Discovering Statistics Using SPSS*, London, SAGE publications, 779 p.

Flury, B. (1997). *A First Course in Multivariate Statistics*. New York, Springer.

Fouche, F.A. & Alberts, N.F. (1977). *Handleiding vir die Negentienveld-Belangstellingsvraelys.* Pretoria, HSRC, 43 p.

Fouche, F.A. & Grobbelaar, P.E. (1983). *Manual for the PHSF Relations Questionnaire*. Pretoria, HSRC, 35 p.

Fouche, F.A. & Verwey, F.A. (1978). *Manual for the Senior Aptitude Tests*, Edition (SAT). Pretoria, HSRC, 82 p.

Foxcroft, C., Paterson, H., le Roux, N. & Herbst, D. (2004). Psychological assessment in South Africa: A needs analysis. The test used patterns and needs of psychological assessment practitioners. Final Report, July 2004. 219 p.

Frey, M.C. & Detterman, D.K. (2003). Scholastic Assessment or g? The Relationship Between the Scholastic Assessment Test and General Cognitive Ability. *Psychological Science*, vol. 15, no. 6, pp 373–378.

Hair, J.R., Anderson, R.E., Tatham, R.L. & Black, W.C. (1998). *Multivariate data analysis. New Jersey,* Prentice-Hall, Inc. 730 p.

Hawkins, D.M. (Ed) (1982). *Topics in Applied Multivariate Analysis*, Cambridge, Cambridge University Press, 362 p.

Hosmer, D.W. Jr. & Lemeshow, S. (2000). *Applied logistic regression,* 2nd ed. New York: Wiley. 375 p.

Houglum, J.E., Aparasu, R.R. & Delfinis, T.M. (2005). Predictors of Academic Success and Failure in a Pharmacy Professional Programme. *American Journal of Pharmaceutical Education,* vol. 69, no. 3, pp. 283-289.

Huberty, C.J. & Olejnik, S. (2006). *Applied MANOVA Discriminant Analysis*. New York, John Wiley & Sons. 479 p.

Johnson, A.J. & Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis.* New Jersey, Prentice Hall, 767 p.

Kass, G.V. (1980). An Exploratory Technique for Investing Large Quantities of Categorical Data. *Applied Statistics*, no. 29, pp. 119-127.

Kass, G.V. (2008). Personal communication.

Kline, R.B. (2004). *Beyond Significance Testing: Reforming data analysis methods in behavioural research.* American Psychological Association, Washington, DC., 323 p.

Kotze, H.N. (1994). *Keuringsmodelle vir universiteitstudierigtings: 'n psigometriese ondersoek.* PhD Thesis, Potchefstroom University for Christian Higher Education. 358 p.

Kutner, H.M., Nachtsheim, C.J., Neter, J. & Li, W. (2005). *Applied Linear Statistical Models* Fifth Ed., McGraw Hill, 1395 p. (15)

Lei, P.-W. & Koehly, L.M. (2000). Linear discriminant analysis versus logistic regression: A comparison of classical errors. Paper presented at the 2000 Annual Meeting of the American Educational Research Association, New Orleans, LA.

Linn, R.L. 1989. Educational Measurement. Macmillan Publishing Company, New York. 610 p.

Liu, R. (2000). Institutional Integration: An Analysis of Tinto's Theory. Paper presented at the 40th Annual AIR Forum, Ohio.

Lourens, A. (2006). *The SAAIR research project on student retention in higher education in South Africa.* Personal communication.

Ludeman, R.B. (2006). *Annual Narrative Report for year one of two year grand period: June 1, 2005 to May 31, 2006*. Kellogg Foundation in Southern Africa Higher Education Retention Data for South Africa (HERD-SA) Grant.

Luo, J. & Jamieson-Drake, D. (2005). Linking Student Precollege Characteristic to College Development Outcomes: The Search of a Meaningful Way to Inform Institutional Practice and Policy. AIR Professional File*. pp.* 1-18.

Macfarlane, D. (2006). Shock varsity dropout stats, *Mail & Guardian*, 22 September 2006.

McLachlan, G.J. (2004). *Discriminant Analysis and Statistical Pattern Recognition,* New York, Wiley-Intersience & Sons, 517 p.

Montgomery, D.C., Peck, E.A. & Vining, G.G. (2001). Introduction to Linear Regression Analysis, Third Ed., New York, John Wiley & Sons. 641 p.

Naes, T. & Mevil, B-H. (2001). Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, 15, 413-426.

Naude, F.P., van Aarde, J.A. & Laubscher, N.F. (1989). Akademiese prestasies van Afrikaans- en Engelssprekende studente tydens tweetalige tersiêre onderrig. *South African Journal of Education,* vol. 9, no. 1, pp. 131-139.

Nunnally, J. & Bernstein, I.H. (1994). *Psychometric theory,* 3rd Ed, New York, McGraw-Hill Inc. 751 p.

NWU (North-West University). (2007). Calendar 2007: Faculty of Arts, Undergraduate programmes. Potchefstroom: Potchefstroom Campus, 167 p.

Owen, K. & Vosloo, H.N. (1999). *Manual for the differential Aptitude tests ,* DAT Form L. Pretoria: HSRC.

Pascarella, E. & Terenzini, P. (1983). Predicting voluntary freshman year persistence/withdrawal behaviour in a residential university: a path analytic validation of Tinto's model. *Journal of Educational Psychology*, vol. 75, no. 2, pp. 215-226.

Pedhazur, E. & Schmelkin, L. (1991). *Measurement, design and analysis*. Hillsdale, NJ: Erlbaum, 819 p.

Peng, C.J., Lee, K.L. & Ingersoll, G.M. (2002a). An Introduction to Logistic Regression Analysis and Reporting. *Journal of Educational Research,* vol. 96, pp. 3-15.

Peng, C.J., So, T.H., Stage, F.K. & John, E.P.St.J. (2002b). The use and interpretation of logistic regression in higher education journals: 1988-1999. *Research in Higher Education,* vol. 43, no. 3, pp. 259-293.

Penny, A.J. (1984). Study habits and academic achievement: a cautionary tale. *South African Journal for Education,* vol. 4, no. 1, pp. 19-23.

Rosenthal, R. & Rosnow, R.L. (1975). The Volunteer Subject. New York, John Wiley & Sons, 266 p.

SAS Institute Inc. (2005a). SAS/STAT, Release 9.1, www.sas.com.

SAS Institute Inc. (2005b). SAS OnlineDoc®, Release 9.1, www.sas.com.

Schepers, J.M. (1990). *Faktorontleding*. Johannesburg, RAU-Drukpers. 186 p.

Schepers, J.M. (1992). *Toetskonstruksie: Teorie en Praktyk*. Johannesburg, RAU-Drukpers. 185 p.

Schepers, J.M. (2004). The power of multiple battery factor analysis in coping with the effects of differential skewness of variables. *SA Journal of Industrial Psychology,* vol. 30, no. 4, pp. 78-81.

Silverman, R.J. (1985). Higher educating as a maturing field? Evidence from referencing practices. *Research in Higher Education*, vol. 23, no. 2, pp. 150-183.

SPSS Inc. (2007). SPSS 16.01.1 for Windows Chicago, www.spss.com

Stevens, J.P. (1992). *Applied multivariate statistics for the social sciences* (2nd edition). Hillside, NJ:Erlbaum.

Steyn, H.S. (jr). (1999). Praktiese beduidendheid: Die gebruik van effekgroottes. *Wetenskaplike Bydraes, Reeks B: Natuurwetenskappe*, nr. 117, Publications Control Committee, PU for CHE, Potchefstroom.

Steyn, H.S. (jr). (2000). Practical significance of the difference in means, *Journal of Industrial Psychology*, vol. 26, no. 3, pp. 1-3.

Steyn, H.S. (jr). (2002). Practically significant relationships between two variables, *SA Journal of Industrial Psychology*, vol. 28, no. 3, pp. 10-15.

Steyn, H.S. (jr) (2006). *Handleiding vir die bepaling van effekgrootte-indekse en praktiese betekenisvolheid*. http://www.puk.ac.za/fakulteite/natuur/skd/indeling.html [Date accessed July 3, 2008].

Tabachnick, B.G. & Fidell, L.S. (2001). *Using Multivariate Statistics*, 4th Ed., Allyn & Bacon, Boston. 966 p.

Thompson, B. (1994). Guidelines for constructors. *Educational and Psychological Measurement*, vol. 54, pp. 837-847.

Tinto, V. (1975). Dropout from higher education: a theoretical synthesis of recent research. *Review of Educational Research*, vol.45, pp. 89-125.

van Eeden, R., de Beer, M. & Coetzee C.H. (2001). Cognitive ability, learning potential, and personality traits as predictors of academic achievement by engineering and other science and technology students. *South African Journal of Higher Education*, vol. 15, no. 1, pp. 171-179.

van der Merwe, R.P. (1999). Psychological assessment in industry, *SA Journal of Industrial Psychology*, vol. 25, no. 3, 8-11.

van Wyk, C.K. (1988). *Die voorspelling van derdevlak-wiskundeprestasie aan 'n universiteit.* PhD Thesis, Potchefstroom University for Christian Higher Education. 241 p.

Venter, J.A. (1995). Die ASAT as voorspeller van akademiese sukses. *South African Journal of Higher Education*, vol. 9, no. 2, pp. 142-147.

Volschenk, P.G. (1997). *Die samestelling van 'n keuringstoets vir eerstejaarstudente aan die PU vir CHO*. MEd Thesis, Potchefstroom University for Christian Higher Education. 241 p.